Máster en Ciencias Actuariales y Financieras

TRABAJO FINAL DE MÁSTER

Título: Modelización del coste en el seguro de automóvil: combinación de regresiones multivariantes

Autoría: David Mateo Argemir

Tutoría: Miguel Angel Santolino Prieto

Curso académico: 2019/2020

UNIVERSITAT DE BARCELONA
Facultat d'Economia i Empresa

Màster de Ciències Actuarials i Financeres

Facultad de Economía y Empresa

Universidad de Barcelona

Trabajo Final de Máster

Máster en Ciencias Actuariales y Financieras

MODELIZACIÓN DEL COSTE EN EL SEGURO DE AUTOMÓVIL: COMBINACIÓN DE REGRESIONES MULTIVARIANTES

Autoría: David Mateo Argemir

Tutoría: Miguel Angel Santolino Prieto

El contenido de este documento es de exclusiva responsabilidad del autor, quien declara que no ha incurrido en plagio y que la totalidad de referencias a otros autores han sido expresadas en el texto.

Modelización del coste en el seguro de automóvil: combinación de regresiones multivariantes

David Mateo Argemir

Resumen

En el presente trabajo de final de máster se utiliza el enfoque general propuesto por Andreas Christmann (2004) en el paper 'An aproach to model complex high-dimensional insurance data'. En este se modelan conjuntos de datos con una estructura de dependencia compleja, con características comunes a las utilizadas en el seguro de automóviles, para construir la prima pura de una cartera de pólizas de automóvil a terceros, utilizando una combinación de diferentes regresiones multivariantes. Concretamente, en el presente trabajo utilizaremos una combinación de regresión Logit multinomial y una Log-Normal.

Palabras clave: Andrea Christmann, Seguro automóvil, prima pura, combinación de regresiones multivariantes, modelo logit multinomial, regresión log-Normal

Abstract

In this master's thesis, I use the general approach proposed by Andreas Christmann (2004) in the paper 'An approach to model complex high-dimension insurance data'. In it, sets of complex dependency structure data are modeled, exploting a database with common characteristics to those used in auto insurance, to construct the pure premium of a third-party auto policies portfolio, using a combination of different multivariate regressions. Specifically, in the present work we use a combination of multinomial logit regression and a Log-Normal.

Keywords: Andrea Christmann, Auto insurance, pure premium, combination of multivariate regressions, multinomial logistic regresion, log-Normal regresion

Índice

1. Introducción

Existen dos ramos en el sector asegurador: vida y no vida. Los seguros del ramo de vida son aquellos que comprenden todos los riesgos que pueden afectar a una persona en su integridad física, salud o existencia, y entre ellos encontramos los seguros de vida, de accidentes personales, de salud o enfermedad y de dependencia. Los seguros del ramo de no vida son aquellos que cubren los riesgos del patrimonio de la persona y de las empresas, y entre ellos encontramos los seguros de automóvil, de ingeniería, multirriesgo, de crédito, de robo, de transportes, de incendios y de responsabilidad civil.

Existen diferencias entre ambos ramos, que se deben de tener en cuenta, como, por ejemplo:

- El componente aleatorio es diferente para ambos ramos. En vida, la variable aleatoria básica es la edad de fallecimiento/supervivencia, mientras que en no vida se trabaja con las variables aleatorias de frecuencia y severidad (o coste).
- Los seguros de no vida se caracterizan por ser seguros a corto plazo (normalmente un año), por lo que el tipo de interés no es demasiado relevante.

En el presente trabajo nos centraremos en el ramo de no vida, y dentro del ramo de no vida, en el seguro de automóviles a terceros. El seguro de automóviles actualmente sigue siendo el más representativo entre los seguros de no vida con un peso en primas del 40,8% en el año 2019 (ICEA (2020). Información del seguro)

Cuando se matricula un vehículo en cualquier país de la UE, hay que asegurarlo de forma obligatoria para cubrir la responsabilidad civil del propietario o conductor (“seguro a terceros”). Este seguro obligatorio es válido en todos los países de la UE y cubre los daños personales o materiales causados a personas distintas del conductor en caso de accidente. Sin embargo, no cubre otros costes, como, por ejemplo, los de reparación del propio vehículo. Si se quiere incluir otras coberturas, se puede contratar un seguro voluntario adicional, como, por ejemplo, a todo riesgo, que incluya otras coberturas como daños corporales al conductor, daños o robo del propio vehículo o de su contenido, actos de vandalismo, asistencia jurídica, etc.

Uno de los principales objetivos de las compañías de seguros es determinar la prima de la póliza que ha de pagar el asegurado (también llamada prima comercial o de tarifa), teniendo en cuenta el riesgo que incorpore la póliza.

Los principios técnicos en que se basa la elaboración de una tarifa constituyen el sistema de tarificación. El objetivo de un sistema de tarificación es que las primas se calculen sean equitativas a cada riesgo, teniendo en cuenta la solvencia y solidaridad entre los asegurados.

Desde la perspectiva actuarial, se diferencian dos sistemas de tarificación (Boj, E; Claramunt, M.M. & Costa, T. (2003). *Matemática Actuarial No Vida. Un Enfoque Práctico*):

- Tarificación a priori o ”Class-rating”: Se establecen primas diferentes por clases. Se denominan “a priori” ya que nos permitirá asignar una prima a una póliza o

riesgo que se incorpora en nuestra cartera sin tener experiencia sobre la siniestralidad de ese riesgo en concreto, excepto la proporcionada en el caso del seguro de automóviles por el Fichero SINCO (Fichero de siniestralidad de conductores, con experiencia de los últimos cinco años). Únicamente conociendo determinadas características de la póliza determinaremos su prima asignándole una siniestralidad esperada. Se recurren a datos de un periodo anterior, relativos a las cuantías de los siniestros, y a las características de los asegurados o tomadores de la póliza.

- Tarificación a posteriori o experience-rating: Podemos entenderla en un sentido estricto, por oposición a la tarificación "a priori", como aquella que presupone la existencia de una prima inicial que se va modificando en base a la experiencia de siniestralidad de aquel individuo (póliza), para dar lugar a las primas de los periodos sucesivos. Da lugar a los sistemas "Bonus-Malus". En un sentido amplio podría entenderse como la actualización de las tarifas mediante la incorporación de nueva información.

La determinación de la prima de tarifa es de fundamental importancia ya que con el importe formado por las primas se afrontarán el pago de los siniestros, por lo que ha de ser lo más precisa posible.

Las técnicas de tarificación en este ramo son bastante variadas, la elección depende de las preferencias y de la información que disponga la compañía. Las compañías utilizan modelos de prima pura o modelos de frecuencia y severidad para estimar la prima. Las ventajas de los modelos de frecuencia y severidad son que se obtiene una mejor comprensión de los factores que afectan a la siniestralidad y permite analizar la frecuencia y la severidad de cada asegurado, por el otro lado, los modelos de prima pura tienen la ventaja que tienen menores requerimientos para el ajuste y hay un menor número de modelos a ajustar y mantener y por lo tanto tienen menores costes de desarrollo y mantenimiento que los modelos de frecuencia y severidad, no obstante, no se puede analizar la frecuencia y severidad por separado y debido a la existencia de un punto masa en el coste igual a 0, debido al alto porcentaje de pólizas sin siniestros, las distribuciones que se pueden ajustar están limitadas.

En este trabajo llevaremos a cabo un proceso de tarificación a priori, mediante modelos de prima pura, en el que la variable aleatoria con la que vamos a trabajar será el coste y nuestro horizonte temporal será de un año.

Modelizaremos el coste y estimaremos la prima pura de un conjunto de pólizas del seguro de automóvil a terceros, utilizando el enfoque propuesto por Andreas Christmann (2004) en el paper 'An approach to model complex high-dimensional insurance data', para modelar el coste agregado por póliza, de una forma alternativa a los métodos tradicionales.

Habitualmente se utilizan los modelos lineales generalizados, para estimar la esperanza del coste. Entre los utilizados destaca el modelo compuesto Tweedie, el cual es más flexible porque no sólo contiene un modelo de regresión para la esperanza de la variable dependiente Y, sino que también tiene un modelo de regresión para la dispersión de Y.

No obstante, en el presente trabajo, modelizaremos el coste agregado de los siniestros por póliza de forma indirecta, troceando la variable aleatoria coste agregado de los siniestros

por póliza, en cuatro intervalos, según diferentes tramos de coste definidos y estimando la esperanza del coste, condicionada a cada uno de los intervalos. Los intervalos son flexibles y no tienen por qué ser cuatro ni los escogidos en el presente trabajo.

En primer lugar, calcularemos las probabilidades de que un individuo pertenezca a alguno de estos cuatro intervalos, condicionado a las características de cada individuo, mediante un modelo de regresión multivariante. A partir de aquí, calcularemos las esperanzas del coste agregado por póliza para cada intervalo definido, por separado, y condicionado a las características de cada individuo, también mediante un modelo de regresión multivariante. Posteriormente combinaremos ambas regresiones multivariantes para obtener la prima pura de cada individuo y observaremos si el enfoque utilizado tiene ventajas prácticas y/o teóricas frente la estimación de la esperanza del coste directamente, que utilizan los modelos tradicionales.

En este trabajo, la probabilidad de que un cliente tenga algún siniestro para los diferentes intervalos de coste, condicionada a sus características (factores de riesgo), se determinará a través del Modelo Logit Multivariante, un tipo de Modelo Lineal Generalizado. Por otro lado, el Modelo de regresión Log-Normal, nos informará de la esperanza del coste de los siniestros, condicionado a las características de cada cliente y condicionado a cada intervalo de coste agregado definido. Para estos cálculos se utilizará el programa estadístico R.

En el capítulo 2 explicaremos la metodología utilizada y describimos el Modelo Logit Multivariante y el Modelo Log-Normal. El capítulo 3 describe las características de los datos utilizados. En el capítulo 4 mostraremos los resultados obtenidos de la aplicación del modelo propuesto para los datos correspondientes a pólizas de automóvil a terceros, contratadas en Francia. El capítulo 5 contiene las conclusiones y debate sobre lo observado en el presente trabajo.

2. Metodología: Tarificación de Seguros de automóvil

Hay que tener en cuenta que el beneficio de la compañía de seguros no está garantizado, ya que el asegurador siempre corre el riesgo de que ocurran siniestros superiores a los previstos, y por lo tanto tener que pagar indemnizaciones que lleven a su balance a la pérdida. Por ello, el asegurador normalmente no acepta riesgos de previsión demasiado difícil o riesgos cuya probabilidad sea desconocida. Y no sólo eso, sino que a la compañía le interesa tener un mix de riesgos en su cartera suficientemente amplio que le permita un buen promedio entre casos sin siniestros y casos con siniestros.

Para ello resulta de vital importancia, un buen cálculo de la prima que ha de pagar el asegurado, tanto para garantizar que las primas serán suficientes para cubrir todos los riesgos de la cartera de la compañía aseguradora, como para garantizar que cada póliza paga por el riesgo que incorpora. En el caso de que una compañía no realice una correcta tarificación para los distintos niveles de riesgo, mientras que sus competidoras sí lo hacen, implicará que perderá a sus mejores riesgos ya que para ellos la competencia les estará ofreciendo mejores precios y se quedarán los peores riesgos los cuales no encontrarán mejores precios en los competidores que lo estén haciendo bien, por ello, la compañía acabará teniendo pérdidas financieras. Por ejemplo, una incorrecta tarificación por arriba es malo desde el punto de vista del tomador de la póliza, ya que la prima estimada será demasiada alta i el cliente tendrá que pagar demasiado dinero en relación con el riesgo que incorpora. Pero también será malo para la compañía de seguros ya que hay el peligro de que el cliente se vaya a otra compañía. Por otro lado, las compañías de seguros evitarán el caso de una incorrecta tarificación por abajo, ya que en este caso tendrían pérdidas. No obstante, puede ser aceptable por parte de la compañía por un periodo corto de tiempo si la compañía está interesada en aumentar su cuota de mercado.

2.1. Prima pura

La prima de tarifa se puede desglosar en la prima pura, más ciertos gastos, como gastos administrativos, impuestos, comisiones y beneficio del asegurador.

En este trabajo nos centraremos en el cálculo de la parte estadística de la prima de tarifa, que es la prima pura. La prima pura es el componente principal de la prima de tarifa, y en términos teóricos, se define como la esperanza matemática de la indemnización que el asegurador se compromete a pagar al asegurado en caso de que se produzca el siniestro, se puede escribir de la siguiente forma,

$$\mathrm{E}(\mathrm{Y} \mid \mathrm{X} = \mathrm{x})^{1}$$

Para construir la prima pura se necesitan estimaciones del coste de los siniestros de una cartera. En nuestros datos dispondremos del coste agregado por póliza, por lo que nos centraremos en modelizar el coste agregado por póliza, para estimar la prima pura para el intervalo de tiempo de un año.

La estimación de la prima pura $\hat{p}'$ (x) debe cumplir los siguientes principios:

- Equidad: que cada asegurado pague según el riesgo que incorpora.

1 $Prima\ pura = E(Y|X = x)$, Esperanza matemática del coste de los siniestros agregados por póliza (Y), condicionada a las características (X) de cada individuo.

- Suficiencia: que, en término esperados, las primas sean suficientes para cubrir todos los riesgos de la cartera de la empresa aseguradora.
- Solidaridad: que, en una cartera homogénea, en la que todos los asegurados tienen el mismo nivel de riesgo, la prima sea igual para todos.

Naturalmente, en nuestros datos no disponemos de la totalidad de la distribución del coste, por este motivo hemos de utilizar modelos de regresión que nos permitan modelar la distribución total y calcular la prima pura para cualquier asegurado. El coste se modelará ajustando distribuciones de probabilidad de carácter continuo:

$$Y = \sum_{i=0}^{N_t} S_i$$

Siendo:

-Y el coste agregado de la cartera
-N_t la variable aleatoria del número de siniestros (frecuencia)
-S_i las severidades de la cartera

La modelización de la prima pura presenta dificultades a la hora de estimarla con los métodos estadísticos clásicos, debido a que:

- La mayoría de los asegurados no tienen siniestros en un año o en un cierto período, por lo que, en la función de distribución del coste, se produce una alta concentración en el punto 0.
- La variabilidad del coste de los siniestros por póliza difiere entre siniestros con coste alto y bajo, ya que entre los siniestros con coste alto podemos encontrar valores muy alejados de la media del grupo de coste alto, como por ejemplo un siniestro de 9 millones de euros, cuando la media del grupo coste alto es 40.000 euros, en cambio entre los siniestros con coste bajo esta variabilidad frente a la media es más reducida.
- Muchas compañías trabajan por módulos (por ejemplo, en el ramo de autos, el módulo de un taller) y esto hace que haya una gran cantidad de siniestros con el mismo importe, por ejemplo 882 euros, creando puntos masa en la función de distribución.
- Los costes extremadamente altos de siniestros son eventos raros, pero contribuyen enormemente a la suma total.
- El coste de los siniestros no siempre es conocido exactamente. Por ejemplo, si un siniestro ocurre en diciembre, el coste exacto del mismo a menudo no será conocido a final de año e incluso a lo mejor ni al final del año siguiente.
- El coste de algunos siniestros son solo estimaciones.
- Existe una compleja estructura de dependencia entre las variables

Si no tuviéramos en cuenta los problemas anteriores, el resultado de la estimación de la prima pura, nos daría un resultado muy sesgado y muy poco preciso.

En el trabajo intentaremos dar solución a los anteriores problemas, para intentar conseguir una prima pura lo más precisa posible, utilizado el enfoque general proporcionado por A. Christmann (2004), que consiste en modelar la variable aleatoria Y del coste agregado de los siniestros por póliza dividiéndola por intervalos de coste. Para ello, definiremos una variable 'C', la cual describirá los diferentes intervalos del coste agregado de los

siniestros por póliza, de cada cliente de nuestra base de datos. Los intervalos que hemos definido y que en la sección de Datos explicaremos, son:

$$C=\begin{cases}\text{Sin Coste}, & \text{si } Y=0\\ \text{Bajo}, & \text{si } Y \in (0,2.000]\\ \text{Medio}, & \text{si } Y \in (2.000,10.000]\\ \text{Alto}, & \text{si } Y > 10.000\end{cases}$$

Entonces, ya no modelizaremos la prima pura de la forma genérica como $E(Y \mid X = x)$ sino que, según el enfoque utilizado en el presente trabajo, vamos a descomponer y modelizar la prima pura respecto a los intervalos de 'C', obteniendo la siguiente expresión:

$$E(Y \mid X = x) = \sum_{c=1}^{4} P(C = c|X = x) \cdot E(Y|C = c, X = x)$$

'C' se define desde 1 hasta 4, correspondiente a las categorías Sin Coste, Bajo, Medio y Alto, y se descompone la esperanza matemática del coste de los siniestros, como la probabilidad de que el individuo pertenezca a una de las 4 categorías 'C' condicionado a las características de cada individuo 'X', multiplicado por la esperanza del coste de los siniestros por póliza condicionado a que pertenezca a alguna de las 4 categorías y condicionado a las características de cada individuo.

Como el primer sumando de la anterior expresión es igual a 0 porque $E(Y|C=0, X=x) \equiv 0$ podemos simplificar la expresión, como:

$$E(Y \mid X = x) = \sum_{c=2}^{3} P(C = c \mid C > 0, X = x) \cdot E(Y|C = c, X = x)$$

Para calcular las probabilidades condicionadas utilizaremos toda la muestra, pero para calcular los costes esperados utilizaremos sólo aquellos que toman valores positivos. Por lo tanto, conseguiremos una reducción del tiempo de cálculo, porque no es necesario calcular $E(Y|X=x;C=0)$. Como a menudo, más del 90% de los clientes no tienen siniestros durante un año, sólo será necesario usar menos del 10% de las observaciones para ajustar las esperanzas condicionales.

Además, este enfoque nos garantiza una mayor flexibilidad, ya que incluso es posible estimar las esperanzas condicionales combinando diferentes modelos de regresión, es decir un modelo para el grupo 'Bajo', un modelo para el grupo 'Medio' y un modelo para el grupo 'Alto'. Esto puede ser de interés especialmente para nuestra clase 'Alto', la cual incluye información del coste de los siniestros a nivel de póliza mayor a 10.000 euros. Esta clase contiene muchas menos observaciones que las otras clases, de modo que podría ser mejor utilizar sólo unas pocas variables explicativas para modelar esta clase, para evitar el ajuste excesivo.

Por otro lado, al utilizar estimadores consistentes para las probabilidades condicionales i las esperanzas condicionales, la prima pura estimada también será consistente, siguiendo el teorema de Slutzky. Aunque este enfoque de estimación podría dar estimaciones sesgadas para conjuntos de datos pequeños, en los que se debería realizar una corrección

del sesgo, en nuestro caso particular, vamos a utilizar una base de datos con 413.170 observaciones de pólizas por lo que será suficiente, para fines prácticos.

2.2. Modelos predictivos

La técnica de tarificación habitual utilizada por las compañías de seguros no vida es aquella basada en el uso de los modelos predictivos.

Estas técnicas se basan en el análisis de los datos actuales recogidos en las bases de datos para poder realizar predicciones sobre futuros sucesos.

Los modelos predictivos explotan patrones de comportamiento identificados en el pasado para poder cuantificar riesgos futuros. Estos modelos capturan la relación entre una serie de variables independientes (factores de riesgo) y la variable a explicar.

Las técnicas predictivas más utilizadas en el campo actuarial son, entre otras:

- Modelos lineales generalizados (GLM): se trata de la técnica de modelización por excelencia en las compañías de seguros no vida. Es una generalización flexible de la regresión lineal que relaciona la distribución aleatoria de la variable dependiente con la parte sistemática a través de la función de enlace.
- Redes neuronales: las redes neuronales artificiales (RNA) son un sistema de interconexión de neuronas que colaboran entre sí para producir un estímulo de salida.
- Árboles de decisión: es un modelo de predicción muy utilizado en el ámbito de la inteligencia artificial. Se trata de una técnica que permite analizar decisiones secuenciales basada en el uso de resultados y probabilidades asociadas.

Para la elaboración de este apartado, se han utilizado los manuales del curso de Econometría Actuarial y Análisis de la supervivencia en seguros (Ayuso.M & Bolancé.C Curso 2019/2020)

En el presente trabajo, nos centraremos en los modelos lineales generalizados (GLM). A continuación, realizaremos una breve introducción de estos y posteriormente explicaremos los dos modelos utilizados en el presente trabajo para obtener dichas estimaciones y poder construir la prima pura.

2.2.1. Modelos Lineales Generalizados

Los modelos lineales generalizados se utilizan como una herramienta que permite valorar y cuantificar la relación existente entre la variable respuesta y las variables explicativas. Se diferencian de los modelos de regresión lineal en tres aspectos:

- Su distribución pertenece a la familia exponencial, conteniendo a la Normal como un caso particular, sin embargo, ya no es necesario que siga una distribución Normal como es el caso de la regresión lineal clásica.
- Su esperanza se relaciona linealmente con las variables explicativas, aunque no directamente, sino a través de una función enlace.

- Su varianza no es necesariamente constante, es una función de su esperanza. Esto es debido a que la variable respuesta sigue una distribución exponencial, siendo habitualmente heterocedástica, lo que hace que hace que su varianza cambie en función de la media.

Estas características hacen que los modelos lineales generalizados sean de gran utilidad en los seguros, pues los datos difícilmente siguen una distribución Normal y tampoco suelen presentar homocedasticidad (implica que la varianza de los errores es constante a lo largo del tiempo).

Un modelo lineal generalizado tiene tres componentes básicos:

- Componente aleatoria: Identifica la variable respuesta y su distribución de probabilidad.
- Componente sistemática: Especifica las variables explicativas (independientes o predictoras) utilizadas en la función predictora lineal.
- Función de enlace: Es una función del valor esperado de Y, E(Y), como una combinación lineal de las variables predictoras.

2.2.1.1. Componente aleatoria

La componente aleatoria de un GLM consiste en una variable aleatoria Y con observaciones independientes (y_1, …, y_N).

En muchas aplicaciones, las observaciones de Y son binarias y se identifican como éxito y fracaso. Aunque de un modo más general, cada Y_i indica el número de éxitos de entre un número fijo de ensayos, y se modeliza como una distribución binomial.

En otras ocasiones cada observación es un recuento, con lo que se puede asignar a Y una distribución de Poisson o una distribución binomial negativa. Finalmente, si las observaciones son continuas se puede asumir para Y una distribución Normal.

Todos estos modelos se pueden incluir dentro de la llamada familia exponencial de distribuciones.

2.2.1.2. Componente sistemática

La componente sistemática de un GLM especifica las variables explicativas, que entran en forma de efectos fijos en un modelo lineal, es decir las variables x_j se relacionan mediante,

$$\alpha + \beta_1 X_1 + \ldots + \beta_k X_k$$

Esta combinación lineal de variables explicativas se denomina predictor lineal.

2.2.1.3. Función de enlace

Tal y como se comentó al inicio de este apartado, la media de la variable respuesta no se relaciona directamente con las variables explicativas, sino mediante una transformación de su esperanza (μ):

$$g(\mu) = x'\beta$$

La transformación viene determinada por la función de enlace g(), siendo está monótona y diferenciable. Las funciones enlace sirven para obligar al modelo a calcular de forma coherente el valor de los parámetros a estimar.

Cada función de distribución, de acuerdo con sus características, tiene asociada una función enlace. En la siguiente tabla podemos ver algunos de las funciones de enlace más utilizadas:

Modelo	**Ligadura**	$\eta_i = g(\mu_i)$
Normal	Identidad	μ_i
Binomial	Logit	$ln\frac{\mu_i}{1-\mu_i}$
Binomial	Probit	$\boldsymbol{\phi}^{-1}(\mu_i)$
Poisson	Log	$\ln \mu_i$
Gamma	Inverse	μ_i^{-1}

Tabla 1: Funciones enlace. *Fuente: Elaboración propia*

2.2.1.4. Estimación de los parámetros de un GLM

Cuando la variable dependiente se distribuye según una distribución de la familia exponencial, los parámetros suelen ser estimados por máxima verosimilitud (MV). La idea es encontrar aquellos valores de los parámetros β_j que hacen máxima la probabilidad de que los datos tratados hayan sido generados por dichos parámetros.

En general la estimación por máxima verosimilitud suele aproximarse por procedimientos iterativos como el Algoritmo de Newton-Raphson o el Método Scoring de Fisher.

En el modelo de regresión lineal los parámetros β_j son los efectos marginales:

$$\frac{\partial y}{\partial xj} = \beta_j,$$

En cambio, en los modelos no lineales (probit, logit, modelos de supervivencia…) los parámetros β_j no coinciden con los efectos marginales.

$$\frac{\partial h(x\beta)}{\partial xj} = \beta_j h'(\mathrm{x}\beta),$$

Donde *h(.)* es la inversa de la función de enlace (por ejemplo: en el modelo Logit es la distribución logística y, por tanto, *h'(.)* es la derivada de *h(.)*. A partir de los parámetros estimados únicamente conocemos el signo de los efectos marginales.

2.2.1.5. Modelos lineales generalizados según tipo de datos

Existen diversidad de modelos que son casos particulares de los modelos lineales generalizados y que dependen de cómo es la variable dependiente.

A continuación, mostramos un cuadro resumen, en el que se pueden observar las diferentes posibilidades en función de la variable dependiente:

Variable dependiente	Definición	Tipo de modelo	Función en R
Cuantitativa continua	Dentro de un rango, puede adquirir infinitos valores numéricos	Modelo de regresión lineal Modelos de supervivencia Modelo Tobit ….	lm()
Categórica dicotómica	Restringida a dos valores ordenados	Elección binaria: logit y probit	glm(family=binomial())
Categórica politómica	Más de dos categorías no ordenadas	Elección multinomial Elección multinomial ordenada Elección multinomial anidada	multinom()
Conteo	Conteo de eventos cuyo universo de probabilidad no es conocido	Modelo de Poisson Modelo Poisson Mixto …	glm(family=log())

Tabla 2: Clasificación de modelos lineales generalizados según la variable dependiente
Fuente: Elaboración propia

En la Tabla 2 anterior observamos, que cuando la variable dependiente es continua, uno de los modelos que puede utilizarse es el modelo de regresión lineal, el cual es un caso particular de los GLM.

Por otro lado, en los casos en los que la variable dependiente es categórica, los modelos que deben utilizarse son el modelo Logit o Probit, si la variable dependiente categórica está restringida a dos valores ordenados. En caso de que la variable dependiente tenga más de dos categorías no ordenadas, uno de los modelos que deberemos usar es el Logit Multivariante.

2.3. Combinación de regresiones multivariantes

Hay muchas posibilidades para estimar las probabilidades de que un individuo pertenezca al grupo ‘Sin Coste’, ‘Bajo’, ‘Medio’ o ‘Alto’ condicionado a las características de cada individuo y las esperanzas del coste agregado por póliza condicionado a cada grupo de ‘C’ y a las características de cada individuo. Algunas combinaciones posibles son:

- Regresión logística multinomial y regresión gamma
- Regresión logística multinomial y regresión vectorial

- Regresión logística kernel y regresión vectorial
- Árboles de clasificación y regresión semi paramétrica
- Redes neuronales
- Regresión logística robusta y árboles de regresión
- Etc.,

En el presente trabajo, para estimar las probabilidades condicionales utilizaremos un Modelo Logit Multivariante, ya que la variable dependiente será categórica politómica, y puede tomar cuatro categorías, 'Sin Coste', 'Bajo', 'Medio' o 'Alto'. Para estimar las esperanzas condicionales del coste de los siniestros por póliza, utilizaremos una regresión Log-Normal, ya que la variable dependiente coste de los siniestros es una variable continua y asimétrica con una larga cola hacia la derecha.

A continuación, describiremos cada uno de los dos modelos utilizados.

2.3.1. Modelo Logit multivariante

En este apartado nos centraremos en el estudio de la modelización de la probabilidad condicionada P(C=c | X=x), es decir, la modelización de la probabilidad de que un individuo este dentro del grupo 'Sin Coste', 'Bajo', 'Medio' o 'Alto', condicionada a las características o factores de riesgo de cada individuo. La suma de la probabilidad de estar en el grupo 'Sin Coste', 'Bajo', 'Medio' o 'Alto' será 1.

Como hemos comentado antes, los Logit Binomiales sirven para modelar una variable categórica dicotómica, es decir solo dos categorías, pero puede interesarnos modelar una salida politómica, es decir, una categórica con más de dos categorías.

En nuestro estudio, trabajaremos con una variable dependiente categórica politómica, ya que 'C' tiene 4 categorías, por lo que, el modelo que vamos a utilizar es el Modelo Logit Multivariante. Este modelo es una técnica analítica que nos permite relacionar funcionalmente una variable politómica con un conjunto de variables independientes. La salida de un modelo Logit multinomial es similar a la de uno binomial, pero como podemos esperar, hay más información, ya que no estamos modelando solamente al evento y=1, sino también y=2, y=3, …, y=k.

2.3.1.1. Ajuste de modelos logit multinomiales con multinom()

Para ajustar el Modelo Logit Multivariante en R, utilizaremos la función multinom(formula, data=datos). El primer término de la función puede ser un factor con dos o más niveles y del lado derecho podemos incluir tantos factores como predictores continuos. Para nuestro trabajo, tomaremos como variable dependiente los grupos de la variable 'C', una categórica politómica con M=4 niveles:

$$C=\begin{cases} \text{Sin Coste}, & \text{si } Y=0 \\ \text{Bajo}, & \text{si } Y \in (0,2.000] \\ \text{Medio}, & \text{si } Y \in (2.000,10.000] \\ \text{Alto}, & \text{si } Y > 10.000 \end{cases}$$

La variable dependiente 'C', la modelaremos como una combinación de 7 variables independientes, que analizaremos con detalle en la sección 3. Datos y que corresponden con: la potencia del vehículo, la antigüedad del vehículo, la edad del conductor, la marca del vehículo, el tipo de combustible del vehículo, la región donde vive el conductor y la densidad de población de la ciudad donde vive el conductor.

Las variables independientes o explicativas antigüedad del vehículo, la edad del conductor y la densidad de población son continuas y las variables potencia del vehículo, marca del vehículo, región y tipo de combustible del vehículo son categóricas, no obstante, no hay problema, ya que en este modelo se acepta cualquier tipo de variable explicativa.

Supongamos que una variable dependiente tiene M categorías. Un valor (normalmente el primero, el último o el valor con la frecuencia más alta) de la variable dependiente se designa como la categoría de referencia. La probabilidad de pertenencia a otras categorías se compara con la probabilidad de pertenencia a la categoría de referencia.

Para la variable dependiente con M categorías, se requiere el cálculo de M-1 ecuaciones, una para cada categoría en relación con la categoría de referencia, para describir la relación entre la variable dependiente y las variables independientes.

En nuestro trabajo, M=4 y la categoría de referencia será el grupo 'Sin Coste', por lo tanto, se requerirá el cálculo de 3 ecuaciones, una para el grupo 'Bajo', otra para 'Medio' y otra para 'Alto', para describir la relación entre la variable dependiente 'C' y las variables independientes y la probabilidad de pertenencia a otras categorías comparadas con la probabilidad de pertenencia a la categoría de referencia.

Para saber si las variables independientes están relacionadas con la variable dependiente se calculan primero los coeficientes β, que son los logit.

No debemos interpretar las β sino su signo. Por ejemplo, en nuestro trabajo, un coeficiente β positivo en la variable explicativa Antigüedad del vehículo, nos estará diciendo que un augmento de esta variable en una unidad, hará augmentar la probabilidad de que aumente la variable dependiente coste 'Bajo' frente la probabilidad de 'Sin Coste', por ejemplo, pero no nos estará dando información de la intensidad de esta relación entre la variable dependiente e independiente.

Para poder explicar la fortaleza de relación entre dos variables, deberemos calcular los Odd-Ratio, que son la $\text{Exp}(\beta_j)$. Los odd-ratio son medidas estandarizadas que permiten comparar el nivel de influencia o fortaleza de las variables independientes sobre la variable dependiente. Las variables independientes están en diferentes escalas, y además están expresadas en logaritmos, y se necesita estandarizar las escalas. La manera de estandarizar y así poder comparar las variables independientes a través de los odd-ratios.

Cuando el $\text{Exp}(\beta_j)>1$ señala que un aumento de la variable independiente, aumenta los odds que ocurra el evento. Cuando el $\text{Exp}(\beta_j)<1$ indica que un aumento de la variable independiente, reduce los odds que ocurra el evento. Cuando más se aleja de 1, más fuerte es la relación entre las dos variables.

La Exp(β) se deben leer como el efecto que tiene el aumento de una unidad en una variable explicativa, en la probabilidad o no de elegir una determinada opción.

Una vez, disponemos de los coeficientes β ya podremos calcular la probabilidad de pertenecer a cada uno de los M categorías. No obstante, cuando hay más de 2 grupos, calcular las probabilidades es un poco más complicado de lo que era en la regresión logística. Para m=2,….,M,

$$P(Y_i = m) = \frac{\exp(Z_{mi})}{1 + \sum_{h=2}^{M} \exp(Z_{hi})}$$

, siendo $Z_{mi} = \alpha_m + \sum_{k=1}^{k} \beta_{mk} X_{ik}$ (en nuestro caso tenemos m=2 para la categoría Bajo, m=3 para la categoría Medio y m=4 para la categoría Alto)

Para la categoría de referencia deberemos calcular,

$$P(Y_i = 1) = \frac{1}{1 + \sum_{h=2}^{M} \exp(Z_{hi})}$$

A tener en cuenta, que cuando M=2 el Modelo Logit Multivariante i la Regresión Logística son lo mismo.

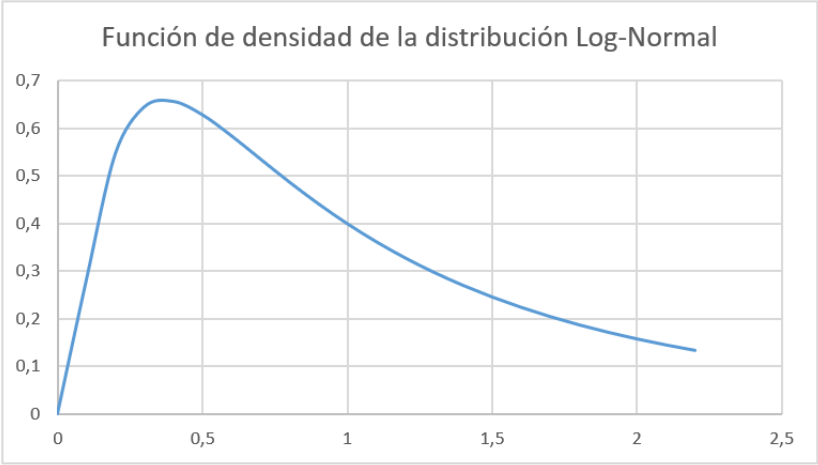
2.3.2. Distribución log-Normal

En este apartado nos centraremos en la modelización de las esperanzas del coste de los siniestros por póliza condicionadas a que pertenezca a cada una de las 4 clases de ‘C’ y a las características o factores de riesgo de cada individuo, es decir, la E(Y| C=c ,X=x)

En realidad como la E(Y| C=0,X=x)=0, sólo deberemos modelizar la E(Y|C=’Bajo’,X=x), E(Y|C=’Medio’,X=x) y la E(Y| C=’Alto’,X=x).

Cuando nos referimos al coste agregado de los siniestros, nos referimos a las pérdidas monetarias derivadas de los siniestros por póliza. En la práctica, para modelar esta variable aleatoria, se utilizan distribuciones continuas que tomen valores positivos.

En nuestro trabajo, vamos a modelizar la variable dependiente Y, de cada clase ‘C’, según una regresión Log-Normal, ya que el coste de los siniestros por póliza suele estar distribuido según una Log-Normal, al ser una variable no negativa (el coste de los siniestros toma valores 0 o valores positivos mayores que 0 en nuestros datos) y asimétrica con una larga cola hacia la derecha (siniestros de coste muy elevado con probabilidad de ocurrencia muy reducida), como la de la distribución Log-Normal. Gráficamente:



Fuente: elaboración propia

2.3.2.1. Regresión Log-Normal con glm()

La variable aleatoria Y, coste de los siniestros por póliza, para cada clase de 'C', tiene una distribución Log-Normal si la variable aleatoria Y∼LnY (logaritmo del coste de los siniestros por póliza) tiene una distribución Normal. Es decir,

$Y \sim LN(\mu, \sigma^2)$, de parámetros μ y σ^2
$ln(Y) \sim N(\mu, \sigma^2)$, de media y varianza

Si los valores del coste no están distribuidos según una Normal, pero tienen una distribución Log-Normal, como es el caso, nos interesará transformarlos en un conjunto de datos distribuidos normalmente, para poder aplicar las técnicas estadísticas habituales, como pruebas de significación y determinación de límites de confianza o predicción.

Para ajustar el modelo glm() en R, primero utilizaremos la relación entre la distribución Log-Normal y la Normal, transformando en primer lugar la variable dependiente coste de los siniestros por póliza, de cada clase, a logaritmos, para transformar los valores a una distribución Normal.

Lo que ajustaremos con este modelo de regresión es la μ=E(ln(Y)). Para cada individuo obtendremos 3 μ, una para cada clase de 'C' ('Baj0', 'Medio' y 'Alto')

$$\mu = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Pero lo que nos interesa modelizar es la E(Y). Como los parámetros μ y σ de la Log-Normal LN (μ,σ^2) se relacionan con los de la distribución Normal N (μ, σ^2), mediante las ecuaciones

$$\mu = Ln(E[Y]) - \frac{1}{2}\sigma^2$$
$$\sigma^2 = Ln(1 + \frac{Var(Y)^2}{E(Y)^2})$$

Vamos a utilizar esta relación para obtener la E(Y) para cada clase de 'C' con la siguiente expresión

$$E(Y) = exp(\mu + \frac{\sigma^2}{2})$$

De este modo, habremos conseguido modelizar las esperanzas del coste de los siniestros por póliza condicionadas a las características o factores de riesgo de cada individuo y a que pertenezca a cada una de las 4 clases de 'C', es decir, la E(Y| C=c ,X=x), mediante una distribución Log-Normal.

En la práctica, podríamos realizar la modelización de E(Y| C=c ,X=x), aplicando una distribución diferente para cada clase de nuestra variable 'C', la cual se ajuste mejor en cada caso, no obstante, en el presente trabajo vamos a aplicar una distribución Log-Normal, para todas las clases de 'C', ya que el objetivo del presente trabajo no es tanto conseguir el mejor ajuste, sino contrastar que si modelizaremos de forma indirecta la suma esperada del importe de los siniestros, a través del enfoque

$$E(Y|X=x) = P(C>0|X=x)\, x \sum_{c=1}^{k} P(C=c|C>0, X=x) \cdot E(Y|C=c, X=x)$$

El ajuste es mejor que el obtenido mediante su estimación directa, como las que utilizan los modelos lineales generalizados.

3. Datos

La base de datos que utilizaremos en este trabajo está disponible en el paquete de R 'CASdatasets', la cual contiene una colección de conjuntos de datos, originalmente para el libro 'Computational Actuarial Science with R' (2015) editado por Arthur Charpentier. El paquete contiene una gran variedad de conjuntos de datos actuariales, y son públicos, por lo que cualquiera puede replicar los resultados de este trabajo.

Para este trabajo utilizaremos los datos correspondientes a 'freMTPLfreq' y 'freMTPLsev'. Los datos de 'freMTPLfreq' contienen una recopilación de los factores de riesgo de 413.169 pólizas del seguro de automóvil a terceros, contratadas en Francia, observadas durante un año. Por otro lado, los datos de 'freMTPLsev' contienen el número de siniestros que ha tenido cada póliza y su correspondiente coste total. Es decir, cada póliza contiene la información agregada del número de siniestros que haya tenido dentro del año de estudio, así como el coste total agregado de los mismos.

3.1. Tratamiento de datos

En la base de datos encontramos 4 variables carácter: potencia del vehículo, marca, tipo de combustible y región (R nos las ha convertido directamente en variables factor "fct") y el resto son variables numéricas: antigüedad del vehículo, edad del conductor y la densidad de habitantes (int o num). Por otro lado, también disponemos del número de póliza, del número de siniestros durante el periodo de exposición de la póliza, la exposición de la póliza y el coste total del siniestro por póliza.

Antes de analizar las variables de nuestra base de datos, vamos a revisar la misma y a tratar los datos para que sean utilizables con los fines del presente trabajo.

3.1.1. Valores missing

En primer lugar, analizando los datos, detectamos que la variable coste total de los siniestros por póliza tiene valores missing, debido a que cuando el coste del siniestro es 0, tiene asignado el valor NA, cuando debería tener 0.

Para poder trabajar con estos datos, vamos a crear una variable nueva para el coste agregado de los siniestros, indicándole que si el número de siniestros de la póliza es 0, entonces, el coste total de los siniestros sea 0, y no NA. Ahora ya podremos trabajar con esta nueva variable.

3.1.2. Exposición

Por otro lado, un problema que nos encontramos es que disponemos del coste del conjunto de siniestros de la póliza, pero cada póliza, tiene una exposición durante el año de análisis, diferente, y se está tratando el coste igual para todas las pólizas (la exposición de cada póliza se distribuye entre 0 y 1, al ser el periodo observado de 1 año). A la hora de realizar predicciones acerca de la cuantía de los siniestros, es fundamental considerar la exposición al riesgo de los asegurados. Esta se puede definir como el periodo de tiempo que estuvo un asegurado expuesto al riesgo.

Por ejemplo, considerando como el año de análisis de Enero a Diciembre y tomando 2 pólizas, una de ellas contratada el 01 de Enero y la otra el 01 de Julio, con el mismo coste total de 200 euros, una tiene una exposición de medio año y la otra de 1 año, por lo tanto, si consideramos que el coste es el mismo en las dos pólizas, la póliza que sólo ha estado expuesta 0,5 años la estamos infravalorando, porque puede ser que en el otro medio año tenga otro siniestro y en cambio en la póliza expuesta 1 año, ya estamos teniendo en cuenta el coste de todo el año.

Para homogeneizar el coste del conjunto de siniestros de todas las pólizas, y utilizarlo en nuestro modelo, crearemos una nueva variable que llamaremos coste total del conjunto de siniestros por póliza corregida y que será el cociente entre la variable coste del conjunto de siniestros por póliza y la exposición, de este modo tendremos una simulación del coste del conjunto de los siniestros por póliza correspondiente a un año.

3.2. Análisis de los factores de riesgo utilizados

Los factores de riesgo son las variables (cuantitativas y cualitativas) que caracterizan a la persona o al bien asegurado y que están relacionadas con la siniestralidad, permitiendo explicarla y predecirla. Las compañías de seguros recolectan cada año mucha información de todos sus asegurados, con el objetivo de conseguir el máximo de factores de riesgo posible, que expliquen la siniestralidad de los asegurados de su cartera.

Las primas de seguros de automóvil son diferentes para cada persona, dependiendo de sus características, ya que, según el principio de equidad, cada asegurado ha de pagar según el riesgo que incorpora. Por ejemplo, dos personas con el mismo historial de siniestralidad pueden tener primas muy diferentes, dependiendo de los factores de riesgo de cada individuo. Si una de las personas tiene unas características que hagan predecir que tiene más riesgo de acabar teniendo peor siniestralidad que el otro, esta persona pagará una prima más cara que el otro.

Algunos factores de riesgo, como su historial crediticio, no tienen nada que ver con su historial de conducción. No obstante, los factores más comunes incluyen la edad, sexo, tipo de vehículo y el uso del vehículo.

La mayoría de las primas de seguro comienzan en un determinado tipo base, y luego se ajustan en función de los factores de riesgo. En la práctica la mayoría de las carteras son heterogéneas, puesto que contienen individuos con diferentes niveles de riesgo.

Nuestra base de datos toma los siguientes 7 factores de riesgo, de cada uno de los asegurados de la base de datos, que detallamos a continuación:

- La potencia del vehículo (agrupada en la siguiente clasificación categórica: d,e,f,g,h,i,j,k,l,m,n,o)
- La antigüedad del vehículo, en años
- La marca del vehículo (agrupada en los siguientes grupos: A (Renault, Nissan i Citroen), B (Volkswagen, Audi, Skoda i Seat), C (Opel, General Motors i Ford), D (Fiat), E (Mercedes, Chrystler i BMW), F (Japonensas (excepto Nissan) y Coreanas) y G (otras))
- El tipo de combustible (Diesel o Gasolina)

- La edad del conductor, en años (en Francia, la edad mínima de conducción es a los 18 años)
- La región en la que vive el conductor, basadas en una clasificación regional francesa (R11, R23, R24, R25, R31, R52, R53, R54, R72, R74)
- La densidad de habitantes (número de habitantes por km2), en la ciudad en la que vive el conductor del vehículo.

A continuación, analizaremos los 7 factores de riesgo o variables explicativas de nuestros datos. En primer lugar, analizaremos las variables cualitativas Potencia, Marca, Combustible y Región y posteriormente las variables cuantitativas Antigüedad del vehículo, Edad del conductor y Densidad de habitantes.

3.2.1. Potencia del vehículo

La primera variable explicativa cualitativa que encontramos, tiene que ver con el vehículo, y es la Potencia del mismo, la cual en los datos nos viene agrupada en 12 categorías (d,e,f,g,h,i,j,k,l,m,n,o). Esta variable está ordenada de forma categórica, no dando más información el conjunto de datos del paquete de R utilizado, sobre el significado de cada categoría.

A continuación, podemos observar cómo se distribuyen los clientes según la variable potencia de su vehículo.

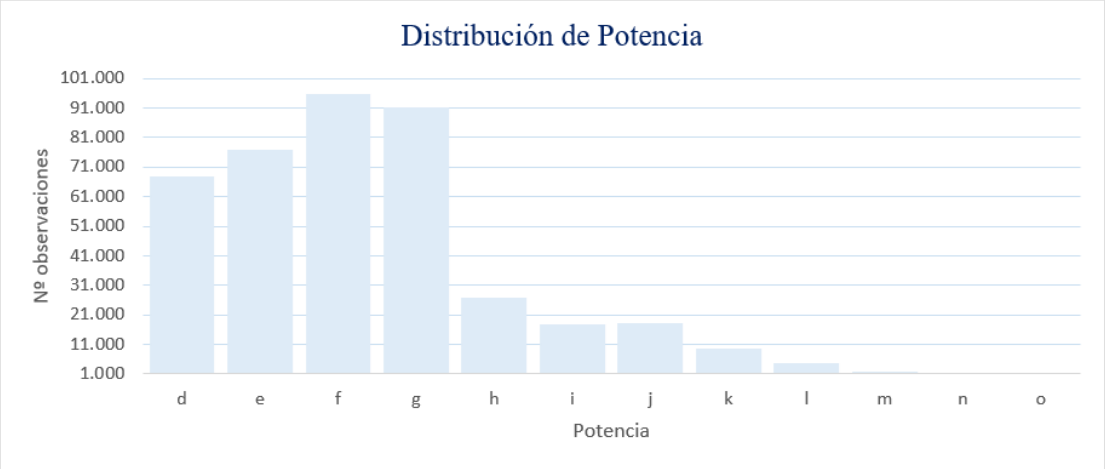


Gráfico 1: Distribución de pólizas según la variable Potencia. *Fuente: Elaboración propia*

La potencia de vehículo que más tienen los clientes de esta base de datos son la categoría ‘f’, representando un 23,2% del total, seguida por la categoría ‘g’, representando un 22,1%, la categoría ‘e’ representando un 18,6% del total y la categoría ‘d’ representando un 16,5%, representando entre las cuatro categorías un 80,3% del total de la muestra. Las categorías ‘m’,’n’ y ‘o’ son las categorías que menos tienen los clientes, con menos del 1% sobre el total.

En la Tabla 1, se analizan las categorías de Potencia del vehículo, en base al coste del conjunto de siniestros por póliza corregido, pudiendo observar como cada categoría tiene un perfil de riesgo diferente entre ellas, en tanto en cuanto, el coste medio y la frecuencia

de siniestros de cada una de las categorías es distinto, así como también su desviación estándar.

Variable	Nº obs.	% obs.	% Coste total	Media	Desviación Estándar	Nº siniestros	%Frecuencia
Potencia	**413.169**	**100,0%**	**100,0%**	**371 €**	**34.930 €**	**16.181**	**3,9%**
d	68.014	16,5%	13,3%	301 €	13.762 €	2.359	3,5%
e	77.022	18,6%	16,2%	322 €	16.756 €	3.201	4,2%
f	95.718	23,2%	37,2%	596 €	62.579 €	3.997	4,2%
g	91.198	22,1%	14,4%	242 €	7.188 €	3.464	3,8%
h	26.698	6,5%	4,4%	254 €	9.213 €	1.000	3,7%
i	17.616	4,3%	8,9%	772 €	70.112 €	722	4,1%
j	18.038	4,4%	2,5%	212 €	3.615 €	710	3,9%
k	9.537	2,3%	1,6%	260 €	8.172 €	380	4,0%
l	4.681	1,1%	0,8%	274 €	6.656 €	162	3,5%
m	1.832	0,4%	0,2%	179 €	2.260 €	76	4,1%
n	1.307	0,3%	0,3%	334 €	8.490 €	56	4,3%
o	1.508	0,4%	0,1%	139 €	2.058 €	54	3,6%

Tabla 1: Descripción estadístico de la variable Potencia, en base al coste total de los siniestros corregido. *Fuente: Elaboración propia*

Hay que tener en cuenta al analizar la variable coste medio por categoría, que, en los datos del coste del conjunto de siniestros por póliza, muchos valores toman valor 0, así como también estamos teniendo en cuenta los valores extremos de cada póliza, y, por lo tanto, es una variable extremadamente sesgada. Para ver si alguna variable está muy impactada por valores extremos, alejados de la media, debemos considerar la desviación estándar, pues valores muy elevados en la desviación estándar, indicaran presencia de valores extremos y que el coste medio esta sesgado.

Según la Tabla 1, podemos observar, como, la última categoría 'o' tiene un coste medio de los siniestros por póliza de 139 euros, el menor coste medio de todas las categorías, así como una frecuencia siniestral del 3,6%, mientras que la primera categoría 'd' tiene un coste medio de 301 euros, mayor que en la mayoría de las categorías y una frecuencia siniestral del 3,5%, por lo que podríamos decir que tienen un comportamiento parecido en cuanto a la frecuencia de declaración de siniestros, pero el coste medio por póliza es mucho mayor en la categoría 'd' que en la 'o'. No obstante, si observamos la desviación estándar de las dos categorías, podemos observar que la de la categoría 'd' es mucho mayor que la de la categoría 'o', por lo que el coste medio por póliza puede estar impactado por valores extremos.

3.2.2. Marca del vehículo

Otra variable que tiene que ver con el vehículo es la Marca. La Marca del vehículo esta agrupada en los siguientes grupos: A (Renault, Nissan i Citroen), B (Volkswagen, Audi, Skoda i Seat), C (Opel, General Motors i Ford), D (Fiat), E (Mercedes, Chrystler i BMW), F (Japonensas (excepto Nissan) y Coreanas) y G (otras).

A continuación, podemos observar cómo se distribuyen los clientes según la variable Marca de su vehículo.

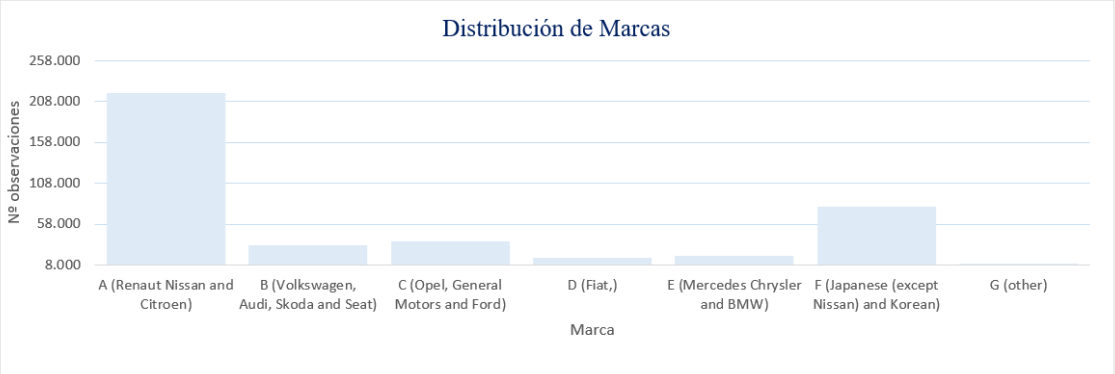


Gráfico 2: Distribución de pólizas según la variable Marca. *Fuente: Elaboración propia*

Como podemos observar en el gráfico, los clientes de nuestra base de datos tienen mayormente un vehículo Renault, Nissan o Citroën (clase A), ya que el 52,8% de las pólizas corresponden a este grupo, después le siguen las marcas japonesas y coreanas (clase F) con un 19,1% del total, representando los dos grupos un 71,9% del total.

Podríamos esperar una equivalencia entre el % de observaciones sobre el total y el % de Coste total de cada grupo, es decir que el coste que aporta cada marca al total sea equivalente al peso que representa cada marca sobre el total, no obstante, como podemos observar en la Tabla 2, las marcas japonesas aportan un coste del 12,5% sobre el total, menor al % que representan sobre el total del 19,1%. Por el contrario, el grupo de Renault, Nissan i Citroën aportan un coste sobre el total del 64,9% mayor al % que representan sobre el total del 52,8%. Por lo que, podríamos concluir que las personas que conducen un vehículo del grupo F, tienen un mejor perfil que los del grupo A.

Variable		Nº obs.	% obs.	% Coste total	Media	Desviación Estándar	Nº siniestros	%Frecuencia
Marca		**413.169**	**100,0%**	**100,0%**	**371 €**	**34.930 €**	**16.181**	**3,9%**
	A (Renaut Nissan and Citroen)	218.200	52,8%	64,9%	456 €	46.795 €	8.942	4,1%
	B (Volkswagen, Audi, Skoda and Seat)	32.638	7,9%	7,2%	337 €	10.554 €	1.470	4,5%
	C (Opel, General Motors and Ford)	37.402	9,1%	7,0%	286 €	10.629 €	1.731	4,6%
	D (Fiat,)	16.723	4,0%	2,5%	228 €	4.788 €	714	4,3%
	E (Mercedes Chrysler and BMW)	19.280	4,7%	4,5%	357 €	10.248 €	833	4,3%
	F (Japanese (except Nissan) and Korean)	79.060	19,1%	12,5%	243 €	14.133 €	2.078	2,6%
	G (other)	9.866	2,4%	1,5%	226 €	4.897 €	413	4,2%

Tabla 2: Descripción estadístico de la variable Marca, en base al coste total de los siniestros corregido. *Fuente: Elaboración propia*

Además, si nos fijamos en la frecuencia siniestral, sí que observamos que el grupo F tiene una ratio del 2,6%, más bajo que la ratio del grupo A del 4,1%.

3.2.3. Combustible

Por otro lado, encontramos otra variable relacionada con el vehículo, como es el Combustible de este. Es decir, si el vehículo asegurado es Diesel o Gasolina.

A continuación, podemos observar cómo se distribuyen los clientes según la variable Combustible de su vehículo.

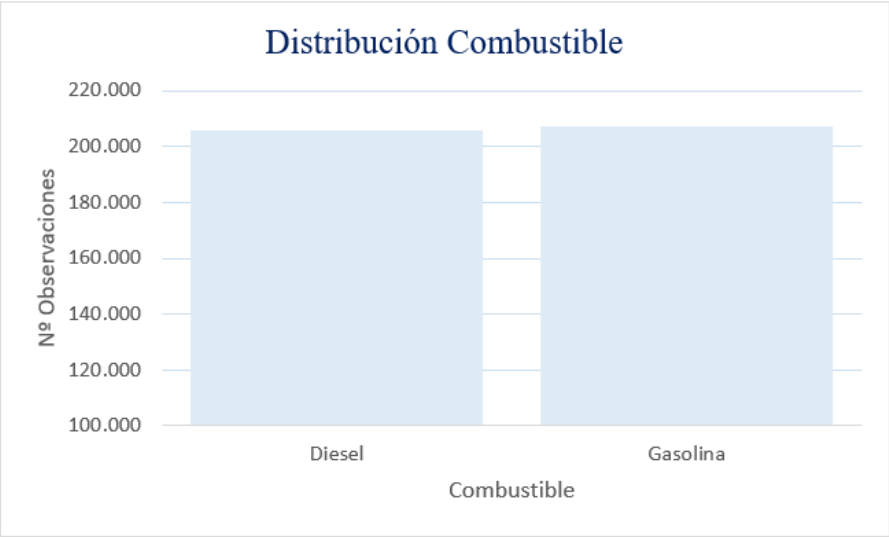


Gráfico 3: Distribución de pólizas según la variable Combustible. *Fuente: Elaboración propia*

Podemos observar en la Tabla 3, como ambos tipos de combustible representan aproximadamente el 50% del total de las pólizas, siendo un poco superior los vehículos gasolina. No obstante, el % de Coste total que aportan las pólizas de gasolina es del 56,9%, superior al % del peso que representan sobre el total del 50,2%.

Variable	Nº obs.	% obs.	% Coste total	Media	Desviación Estándar	Nº siniestros	%Frecuencia
Combustible	**413.169**	**100,0%**	**100,0%**	**371 €**	**34.930 €**	**16.181**	**3,9%**
Diesel	205.945	49,8%	43,1%	321 €	15.378 €	8.446 €	4,1%
Gasolina	207.224	50,2%	56,9%	421 €	46.880 €	7.735 €	3,7%

Tabla 3: Descripción estadístico de la variable Combustible, en base al coste total de los siniestros corregido. *Fuente: Elaboración propia*

Si observamos el coste medio de los vehículos de Gasolina de la Tabla 3, observamos que el coste medio es más elevado en los vehículos de Gasolina, no obstante, la desviación estándar de los vehículos Gasolina es mucho mayor que la de los vehículos Diesel, por lo que la variable Coste total puede estar impactada por siniestros con valores extremos. Si nos fijamos en la frecuencia siniestral de los vehículos Gasolina es menor que la de los vehículos Diesel, siendo 3,7% vs 4,1%.

3.2.4. Región

Ahora nos encontramos ante una variable que corresponde a información demográfica del asegurado, como es la Región en la que vive el conductor, basadas en una clasificación regional francesa (R11, R23, R24, R25, R31, R52, R53, R54, R72, R74).

A continuación, podemos observar cómo se distribuyen los clientes según la variable Región de su vehículo.

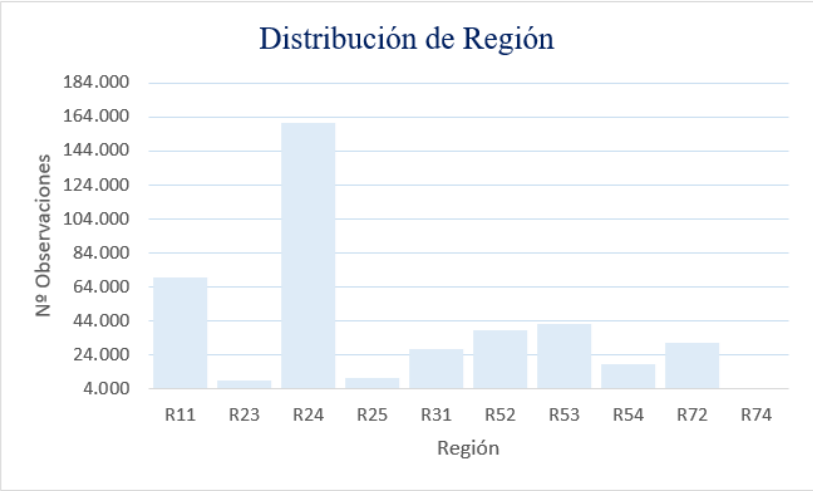


Gráfico 4: Distribución de pólizas según la variable Región. *Fuente: Elaboración propia*

Podemos observar en la Tabla 4, como la región donde viven más asegurados de esta base de datos, es la región R24, representando un 38,9% del total de las pólizas y siendo la que más coste total representa sobre el total, aportando un 41,1% del coste total. La segunda región con más peso es la R11, representando un 16,9% del total de las pólizas y aportando un 23,2% del coste total. Conjuntamente, representan un 55,8% del total de las pólizas, y aportan un 64,3% del coste total.

Variable	Nº obs.	% obs.	% Coste total	Media	Desviación Estándar	Nº siniestros	%Frecuencia
Región	**413.169**	**100,0%**	**100,0%**	**371 €**	**34.930 €**	**16.181**	**3,9%**
R11	69.791	16,9%	23,2%	510 €	69.751 €	2.591	3,7%
R23	8.784	2,1%	0,6%	113 €	1.269 €	220	2,5%
R24	160.601	38,9%	41,1%	392 €	27.129 €	6.475	4,0%
R25	10.893	2,6%	2,4%	333 €	19.426 €	452	4,1%
R31	27.285	6,6%	7,1%	397 €	20.344 €	944	3,5%
R52	38.751	9,4%	9,4%	371 €	18.589 €	1.576	4,1%
R53	42.122	10,2%	8,4%	307 €	19.485 €	1.871	4,4%
R54	19.046	4,6%	2,9%	232 €	5.741 €	800	4,2%
R72	31.329	7,6%	4,4%	217 €	5.395 €	1.055	3,4%
R74	4.567	1,1%	0,5%	165 €	2.580 €	197	4,3%

Tabla 4: Descripción estadístico de la variable Región, en base al coste total de los siniestros corregido. *Fuente: Elaboración propia*

La región R23 es la que tiene una frecuencia siniestral más baja, con un 2,5% de siniestros sobre las pólizas de esta región, así como también tiene el coste medio por póliza más bajo.

3.2.5. Antigüedad del vehículo

A continuación, nos encontramos con la primera variable cuantitativa y la última variable relacionada con el vehículo, como es la Antigüedad de este.

Para poder realizar un mejor análisis, se ha agrupado la variable Antigüedad del vehículo en 6 grupos de años: vehículos nuevos, de 1 año de antigüedad, de 2 años, de 3 a 5 años, de 6 a 10 años y de más de 10 años de antigüedad.

A continuación, podemos observar cómo se distribuyen los clientes según la variable Antigüedad de vehículo.

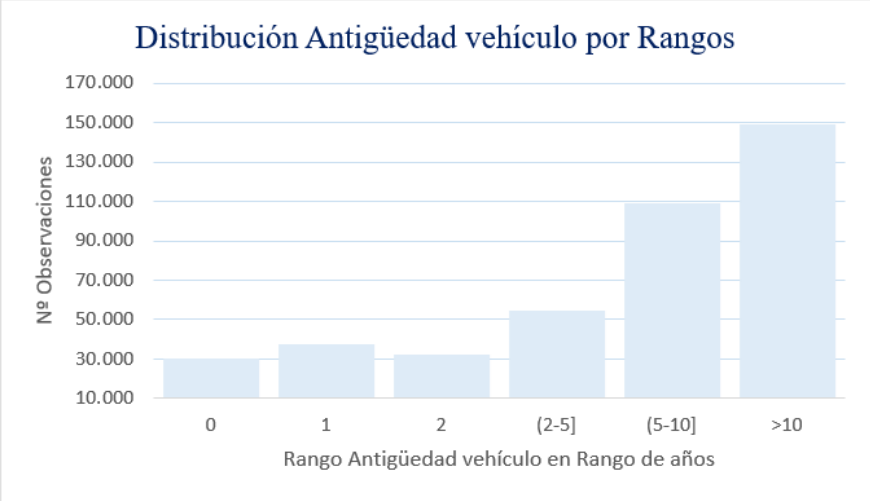


Gráfico 5: Distribución de pólizas según la variable Antigüedad del vehículo. *Fuente: Elaboración propia*

Como podemos observar en el Gráfico 6, 'Frecuencia del conjunto de siniestros por póliza y Rangos de Antigüedad del vehículo' los vehículos nuevos (0 años), son los que presentan una menor frecuencia de siniestros, posiblemente debido a los nuevos sistemas de seguridad activa de los vehículos nuevos, y la misma va aumentando conforme aumentan los años del vehículo.

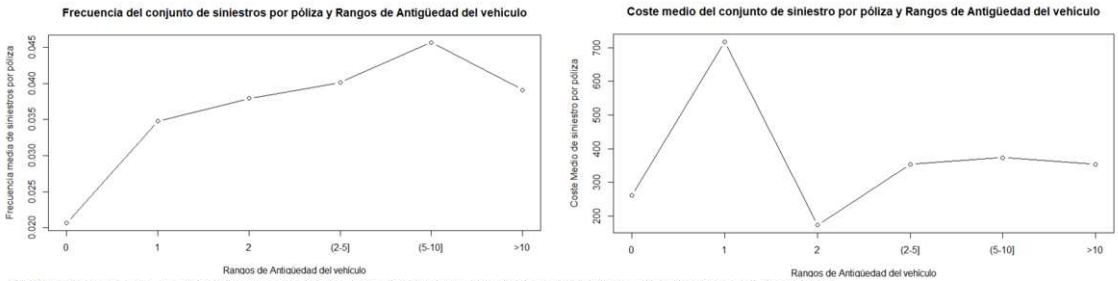


Gráfico 6: Frecuencia de siniestros y coste medio por póliza y por Antigüedad del Vehículo. *Fuente: Elaboración propia*

Parece entonces que hay una relación positiva entre la antigüedad del vehículo y la declaración de siniestros, es decir, a más antigüedad, mayor frecuencia de siniestros.

Por lo que también vamos a estudiar la fuerza de dicha relación, mediante el coeficiente de correlación de Spearman entre la antigüedad del vehículo y el número de siniestros.

Los valores que puede tomar el coeficiente de correlación siempre se encuentran comprendidos entre $-1 \leq r \leq +1$. El resultado es de 0.012 por lo que podemos afirmar, por el signo del resultado, que la relación es positiva entre ambas variables, aunque el valor está muy próximo a 0 por lo que también podemos concluir que no existe una asociación por rangos entre ambas variables. No obstante, puede que exista una relación entre ellas más compleja.

Por lo que respecta al coste medio de los siniestros por póliza, parece que no existe la misma tendencia que sigue la tendencia, sino que existen siniestros con valores extremos en la antigüedad de vehículo del segundo año que hacen que no quede muy clara si hay una relación entre la antigüedad del vehículo y el coste medio.

3.2.6. Edad del conductor

A continuación, analizaremos otra variable cuantitativa que constaría dentro del grupo de información demográfica, como es la Edad del conductor.

En este caso, también se ha realizado una agrupación en 5 grupos de edad: de 18 a 30 años, de 31 a 45 años, de 45 a 50 años, de 55 a 65 años y mayores de 65 años.

A continuación, podemos observar cómo se distribuyen los clientes según la variable Edad del Conductor por rangos.

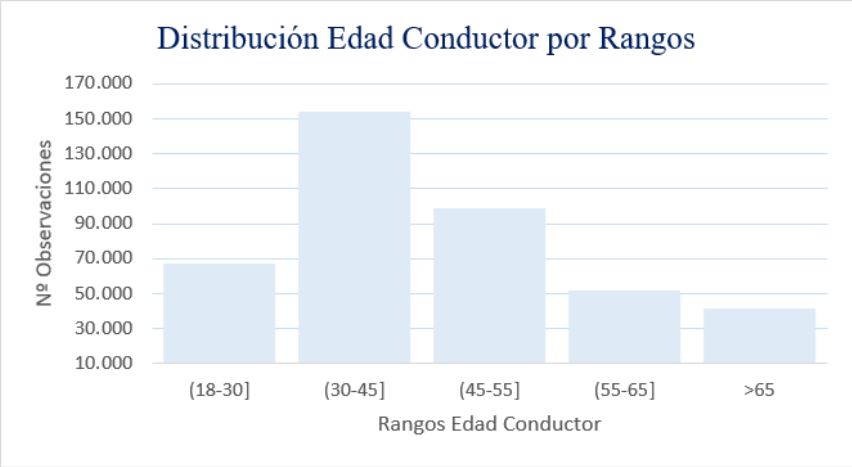


Gráfico 7: Distribución de pólizas según la variable Edad del Conductor. *Fuente: Elaboración propia*

Podemos observar, en el Gráfico 8, como los conductores más jóvenes, de entre 18 y 30 años, son los que tienen una frecuencia de siniestros más alta, y de mayor coste medio. Por otro lado, también se observa una punta de frecuencia en conductores entre 46 y 55 años, lo cual podría ser explicado por el hecho de que este grupo de personas suele ser el que tiene los hijos con edad por empezar a conducir y conducen como segundos conductores, teniendo como hemos visto anteriormente este grupo mayor frecuencia siniestral. Por otro lado, también existe una punta de frecuencia en los mayores de 65 años.

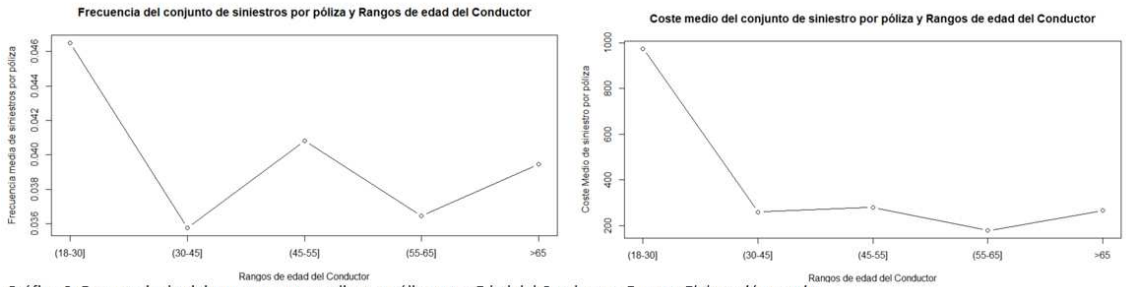


Gráfico 8: Frecuencia de siniestros y coste medio por póliza y por Edad del Conductor. *Fuente: Elaboración propia*

El resultado del coeficiente de correlación de Spearman es de -0.005 por lo que podemos afirmar, por el signo del resultado, que la relación es inversa entre la edad del conductor y el número de siniestros, es decir que a medida que aumenta la edad, disminuye la frecuencia de siniestros, aunque el valor está muy próximo a 0 por lo que también podemos concluir que no existe una relación por rangos entre ambas variables.

3.2.7. Densidad de habitantes

Por último, veremos la última variable explicativa cuantitativa, la cual también se encuentra englobada dentro del grupo de información demográfica, como es la Densidad de habitantes, que corresponde al número de habitantes por km2, en la ciudad en la que vive el conductor del vehículo.

Para ello, también hemos realizado una agrupación correspondiente a 5 grupos de Densidad de habitantes: menos de 200 habitantes, entre 200 y 500, entre 500 y 1000, entre 1000 y 2000 y de más de 2000 habitantes.

A continuación, podemos observar cómo se distribuyen los clientes según la variable Densidad de habitantes por km2 por rangos.

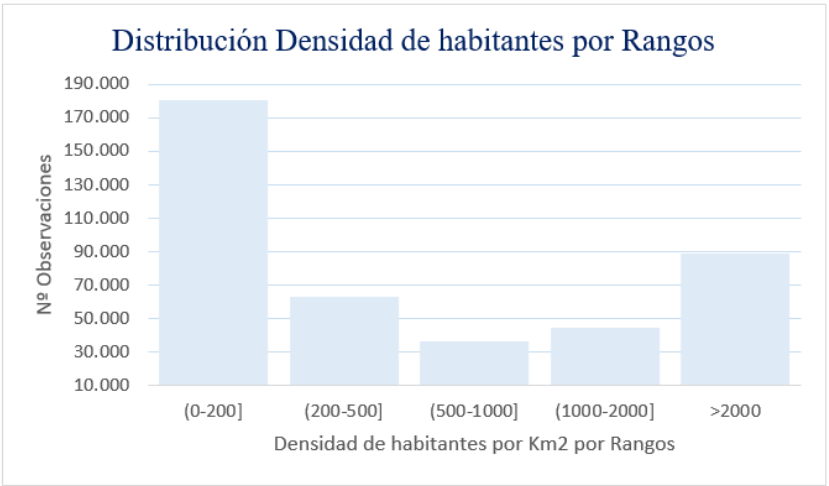


Gráfico 9: Distribución de pólizas según la variable Densidad de habitantes. *Fuente: Elaboración propia*

Como podemos observar en el Gráfico 10, parece que en poblaciones con densidades de habitantes por km2 más pequeñas, la frecuencia de siniestros es menor y que aumenta a medida que aumenta la densidad de población hasta el tramo de densidad de 1000 a 2000 habitantes por km2, disminuyendo a partir de poblaciones con densidad de habitantes

superior a 2.000 habitantes por km2, hasta la altura de poblaciones entre 500 y 1.000 habitantes por km2.

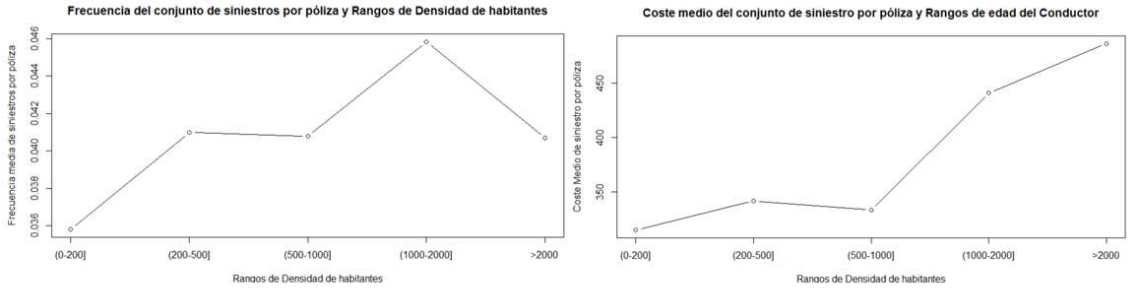


Gráfico 10: Frecuencia de siniestros y coste medio por póliza y por Densidad de Habitantes. Fuente: Elaboración propia

El resultado del coeficiente de correlación de Spearman es de 0.012, por lo que podemos afirmar, por el signo del resultado, que la relación es positiva entre la densidad de habitantes y el número de siniestros, es decir que a medida que aumenta la densidad de habitantes, aumenta la frecuencia de siniestros, aunque el valor está muy próximo a 0 por lo que también podemos concluir que no existe una relación por rangos entre ambas variables.

3.3. Análisis de la distribución del número y coste de los siniestros por póliza

Ahora vamos a analizar la distribución del número de siniestros y el coste de estos. Hay 16.181 siniestros en total. El 96,3% de los clientes no tubo siniestros, un 3,5% tubo un siniestro y 0,2% tubo dos siniestros. El resto tubo tres o cuatro siniestros.

Para aplicar el enfoque de nuestro trabajo, se ha dividido la variable coste del conjunto de siniestros por póliza en 4 clases que serán los cuales conformarán nuestra variable 'C'. Las 4 clases se han dividido con el objetivo de que haya cantidad suficiente de datos en cada grupo, para poder aplicar modelos estadísticos sobre ellos. El corte más difícil corresponde a la clase 'Alto' ya que es en el que se encuentran mayor variabilidad del coste de los siniestros y en el que hay menos datos. Hemos considerado que el punto de corte para la clase 'Alto' debía ser siniestros mayores a 10.000 euros para poder tener un volumen de observaciones de por lo menos 1.000 observaciones, ya que si ampliábamos el punto de corte obteníamos muy pocas observaciones y por lo tanto esta clase era muy poco consistente. Existe una gran literatura para modelar valores extremos, así como para determinar el punto de corte, por ejemplo, el paper 'Application of the Peaks-Over-Threshold Method on Insurance Data' de Max Rydman.

La Tabla 5 muestra las 4 clases, y en ella podemos observar que aproximadamente el 96% de las pólizas no tuvieron ningún siniestro en el periodo de estudio, a este segmento le hemos asignado el nombre 'Sin Coste'. Aproximadamente un 2,3% de las pólizas tuvieron siniestros con un coste menor a los 2.000 euros, a este segmento le llamaremos coste 'Bajo', un 1,1% de las pólizas tuvieron siniestros de un importe total entre 2.000 euros y 10.000 euros, a este segmento le asignaremos el nombre de coste 'Medio', y podemos observar como la suma del importe de sus siniestros contribuye en un 12,1% a la suma del coste total de la cartera. También es interesante observar como sólo el 0,3% de todas las pólizas tubo siniestros con un importe mayor a 10.000 euros, que

denominaremos como ‘Alto’, pero que la suma del coste total de este segmento contribuye en un 81,6% a la suma del coste total de la cartera. La póliza con el mayor coste total de siniestros asciende a 18.246.700 euros.

Importe siniestro	Nº obs.	% obs.	% Coste total	Media	Mediana	Desviación Estándar
Sin Coste	397.779	96,3%	0,0%	0 €	0 €	0 €
Bajo	9.531	2,3%	6,3%	1.006 €	1.150 €	521 €
Medio	4.565	1,1%	12,1%	4.075 €	3.387 €	1.976 €
Alto	1.294	0,3%	81,6%	96.716 €	21.249 €	616.833 €

Tabla 5: Descripción del coste de los siniestros por tramos. *Fuente: Elaboración propia*

La media total del coste de los siniestros, sin tener en cuenta ninguna variable explicativa es de 371 euros. No obstante, hay que tener en cuenta que la distribución de estos valores tiene muchos valores que toman 0, así como existen siniestros con valores extremos y, por lo tanto, es extremadamente sesgada.

Si representamos gráficamente el coste total de los siniestros en un box plot, en el Gráfico 5, podemos observar como la mayoría de los datos se encuentran en el valor 0 y como los bigotes se van extendiendo hasta el valor máximo de 18.246.700.

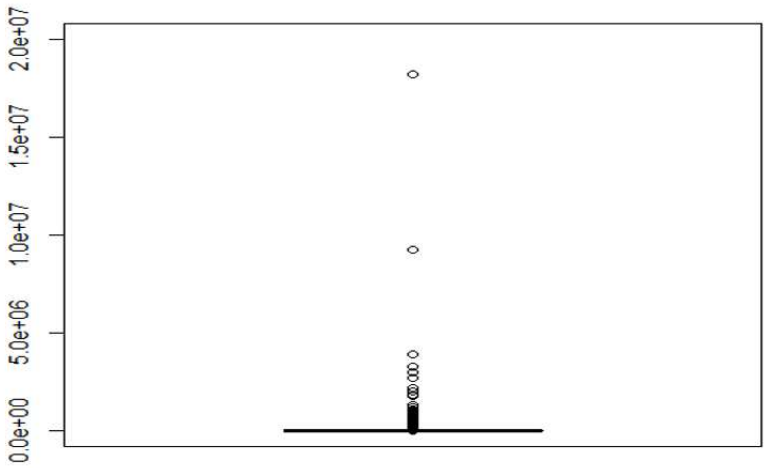


Gráfico 11: Bloxplot variable coste de los siniestros por póliza, corregida. Fuente: Elaboración propia

Queda claro, de la Tabla 5, que las pólizas con siniestros por encima de 10.000 euros contienen información muy valiosa concerniente a la prima pura, aunque el nombre de pólizas con siniestros extremos son muy pocas.

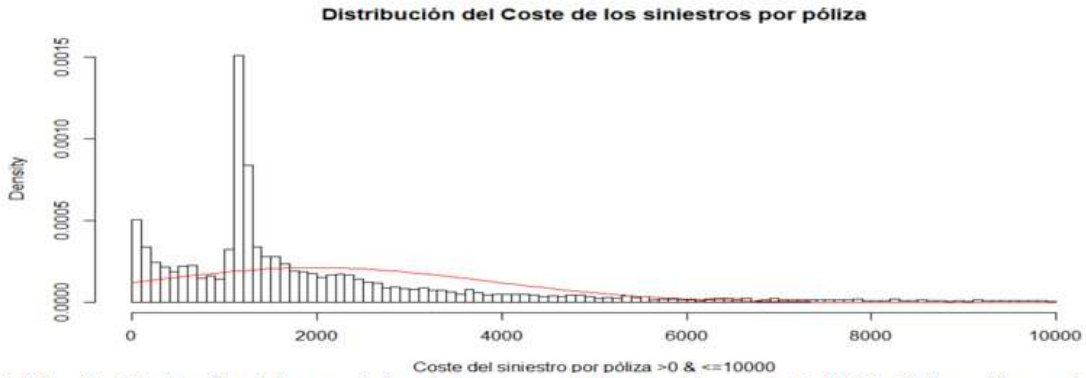


Gráfico 12: Distribución del coste de los siniestros por póliza en el intervalo (0-1000]. *Elaboración propia*

En el Gráfico 12 podemos observar cómo se distribuye la variable coste de los siniestros por póliza en el intervalo entre 0 y 10.000 euros, pudiendo observar como existen puntos masa alrededor de los 1.000 euros, que se corresponden a los módulos con los que trabajan las compañías de seguro (por ejemplo, en el ramo de autos, el módulo de un taller).

En nuestro trabajo, vamos a modelizar esta variable mediante una distribución Log-Normal, al tratarse de una variable no negativa y asimétrica con una cola larga hacia la derecha, no obstante, como ya hemos comentado anteriormente, podríamos ajustar una distribución específica para el tramo que agrupa los puntos masa o para el tramo de pólizas con siniestros de más de 10.000 euros, para conseguir ajustar de la forma más precisa esta variable.

4. Resultados

En la siguiente sección se detallarán los resultados obtenidos tras la construcción del modelo de regresión Logit Mutlivariante para obtener las probabilidades de que un asegurado *j* este dentro de la clase Sin Coste, Bajo, Medio o Alto, así como los resultados obtenidos tras ajustar la regresión Log-Normal para calcular la esperanza del coste de cada clase de 'C' para cada asegurado.

Indicar que antes de aplicar los modelos de regresión, se ha realizado un transformación de la variable Densidad de habitantes, dividiéndola por 1.000 con el objetivo que los coeficientes de esta variable tuvieran sentido con el resto de variables, ya que al ser una variable en miles de habitantes sus coeficientes quedaban muy reducidos.

Posteriormente se estimará la prima pura para cada individuo *j* de nuestra base de datos y se compararán los resultados obtenidos con los costes reales que tuvo cada individuo, para comprobar si el modelo está ajustando bien o no. También se compararán los resultados obtenidos, con los que se hubieran obtenido al modelar la prima pura teniendo en cuenta sólo dos grupos Sin Coste y Con Coste para comprobar que al trocear nuestro modelo en 4 intervalos se ajusta mejor al coste real de las pólizas que sólo diferenciando si ha tenido o no siniestros, ya que los parámetros irán variando entre los intervalos dando más margen, que si lo modelizamos directamente.

Nuestro objetivo será demostrar que, modelizando la prima pura de una forma indirecta, según los 4 tramos definidos, se ajusta mejor y tiene ventajas sobre la estimación directa de esta cantidad.

4.1. Resultados modelo de regresión Logit Multivariante

En esta sección mostraremos los resultados de la modelización de la probabilidad de que cada individuo pertenezca a cada una de las clases de C, condicionada a las características de cada individuo, $P(C = c|X = x)$, en el que nuestra variable dependiente son los grupos de la variable 'C', una categórica politómica con M=4 niveles:

$$C=\begin{cases}\text{Sin Coste}, & \text{si Y=0}\\ \text{Bajo}, & \text{si Y} \in (0,2.000]\\ \text{Medio}, & \text{si Y} \in (2.000,10.000]\\ \text{Alto}, & \text{si Y} > 10.000\end{cases}$$

Y nuestras variables independientes son la potencia del vehículo, la antigüedad del vehículo, la edad del conductor, la marca del vehículo, el tipo de combustible del vehículo, la región donde vive el conductor y la densidad de población de la ciudad donde vive el conductor

A continuación, mostramos los coeficientes (β_j) obtenidos por el modelo de regresión Logit Multivariante:

		Bajo		Medio		Alto	
		Coeficientes	**P value**	**Coeficientes**	**P value**	**Coeficientes**	**P value**
Referencia		-3,894	0,000	-3,903	0,000	-4,871	0,000
Potencia Vehículo							
	e	0,090	0,016**	0,104	0,048**	0,053	0,595
	f	0,118	0,001***	0,096	0,063*	0,082	0,397
	g	0,046	0,205	0,085	0,096*	0,049	0,611
	h	0,081	0,118	0,094	0,190	-0,015	0,914
	i	0,160	0,005***	0,150	0,063*	0,172	0,249
	j	0,205	0,000***	0,077	0,355	0,280	0,056*
	k	0,188	0,012**	0,252	0,012**	-0,100	0,649
	l	0,171	0,123	-0,150	0,368	0,405	0,097*
	m	0,079	0,623	0,228	0,288	0,020	0,963
	n	0,116	0,525	0,218	0,382	-0,289	0,622
	o	0,275	0,137	0,076	0,767	0,048	0,924
Antigüedad Vehículo		-0,010	0,000***	-0,001	0,769	-0,008	0,176
Edad del conductor		0,005	0,000***	-0,011	0,000***	-0,020	0,000***
Marca							
	Japanese (except Nissan) or Korean	-0,838	0,000***	-0,293	0,000***	-0,271	0,102
	Mercedes, Chrysler or BMW	-0,069	0,326	-0,060	0,556	0,287	0,129
	Opel, General Motors or Ford	0,086	0,141	0,029	0,736	0,170	0,309
	other	-0,057	0,481	-0,062	0,604	0,028	0,906
	Renault, Nissan or Citroen	-0,044	0,389	-0,083	0,262	0,094	0,526
	Volkswagen, Audi, Skoda or Seat	0,016	0,792	-0,008	0,931	0,138	0,415
Combustible							
	Regular	-0,063	0,006***	-0,067	0,039**	-0,099	0,104
Región							
	R23	-0,658	0,000***	-0,190	0,100*	-0,156	0,476
	R24	0,117	0,006***	-0,134	0,015**	-0,053	0,615
	R25	0,209	0,003***	-0,118	0,263	-0,076	0,703
	R31	-0,142	0,016**	-0,074	0,305	0,010	0,942
	R52	0,162	0,001***	-0,130	0,057*	-0,046	0,723
	R53	0,234	0,000***	-0,071	0,290	0,012	0,923
	R54	0,175	0,004***	0,002	0,981	-0,029	0,858
	R72	-0,137	0,016**	-0,067	0,345	0,048	0,719
	R74	0,278	0,007***	0,160	0,246	-0,131	0,666
Densidad de habitantes		0,013	0,000***	0,019	0,000***	0,009	0,190

Tabla 6: Resultados del modelo logit multivariante, del cual la variable dependiente es la categoría del coste total de los siniestros por póliza 'Sin Coste', 'Bajo', 'Medio' o 'Alto'. Los asteriscos indican los coeficientes significativos al 1% (***), al 5% (**) y al 10%(*) en el test de significación individual. *Fuente: Elaboración Propia*

En los resultados anteriores se obtienen los coeficientes y los p valores. Cada variable independiente y la de referencia aparecen 3 veces, una para cada una de las categorías de 'C' denominadas 'Bajo', 'Medio' y 'Alto'. La categoría 'Sin Coste' no aparece ya que es la categoría de referencia. Los resultados obtenidos, describen la relación entre la variable dependiente 'Bajo', 'Medio' y 'Alto' con las variables independientes y la probabilidad de pertenencia a estas categorías comparadas con la probabilidad de pertenencia a la categoría de referencia 'Sin Coste'. Correspondiendo a tres ecuaciones:

$$ln(\frac{P(C=Bajo)}{P(C=Sin\ Coste)}) = b_{10} + b_{11} \cdot Potencia + b_{12} \cdot Antigüedad\ vehículo + b_{13} \cdot Edad\ conductor + b_{14} \cdot Marca + b_{15} \cdot Combustible + b_{16} \cdot Región + b_{17} \cdot Densidad$$

$$ln(\frac{P(C=Medio)}{P(C=Sin\ Coste)}) = b_{20} + b_{21} \cdot Potencia + b_{22} \cdot Antigüedad\ vehículo + b_{23} \cdot Edad\ conductor + b_{24} \cdot Marca + b_{25} \cdot Combustible + b_{26} \cdot Región + b_{27} \cdot Densidad$$

$$ln(\frac{P(C=Alto)}{P(C=Sin\ Coste)}) = b_{30} + b_{31} \cdot Potencia + b_{32} \cdot Antigüedad\ vehículo + b_{33} \cdot Edad\ conductor + b_{34} \cdot Marca + b_{35} \cdot Combustible + b_{36} \cdot Región + b_{37} \cdot Densidad$$

donde los b_{ij} son los coeficientes de regresión del modelo.

Como es un modelo no lineal, los coeficientes no son los efectos marginales, así pues, no podemos interpretar la estimación de los coeficientes de forma directa sobre los valores de las variables, aunque sí podemos sacar conclusiones por sus signos. De forma que, si el coeficiente es positivo, provoca que la probabilidad de que un cliente esté dentro de la clase 'Bajo', 'Medio' o 'Alto frente a la 'Sin Coste' se incrementa. Por el contrario, en los casos en los que β<0 la probabilidad de que una cliente esté dentro de la clase 'Bajo', 'Medio' o 'Alto frente a la 'Sin Coste' disminuye. Por ejemplo, por cada aumento en una unidad de la variable edad del asegurado, el logaritmo del ratio de las dos probabilidades, *P(C=Bajo)/P(C=Sin Coste),* se incrementa en 0,005, el logaritmo del ratio de las probabilidades, *P(C=Medio)/P(C=Sin Coste),* disminuye en -0,011 y el logaritmo del ratio de las probabilidades *P(C=Alto)/P(C=Sin Coste),* disminuye en -0,020. Por lo tanto, en general, cuanto mayor sea un asegurado tendrá más preferencia por tener un coste por póliza de clase Bajo, que, por un coste por póliza de la clase Sin Coste, y menor preferencia por tener un coste por póliza de la clase Medio o Alto, que por un coste por póliza de la clase Sin Coste. Por otro lado, con la variable Antigüedad del vehículo observamos que, por cada aumento en una unidad de la variable Antigüedad del vehículo, el logaritmo del ratio de las dos probabilidades, *P(C=Bajo)/P(C=Sin Coste),* disminuye en -0,010, el logaritmo del ratio de las probabilidades, *P(C=Medio)/P(C=Sin Coste),* disminuye en -0,001 y el logaritmo del ratio de las probabilidades *P(C=Alto)/P(C=Sin Coste),* disminuye en -0,008. Por lo tanto, en general, cuanto mayor sea a Antigüedad del vehículo el cliente tendrá más preferencia por tener un coste por póliza de la clase Sin Coste, que por las clases Bajo, Medio o Alto.

También hemos de tener en cuenta, que, de cada variable explicativa, hay una categoría en la que no aparecen los coeficientes, ya que es la que se toma de referencia y sus resultados aparecen agrupadas en el primer coeficiente de referencia.

Por otro lado, también debemos analizar la significación de los coeficientes. Esta significación viene determinada por el valor p (p value en inglés). Que una asociación entre dos variables sea estadísticamente significativa quiere decir que puede descartarse que haya aparecido por azar, y por lo tanto si se mantienen las otras variables constantes la variable con coeficiente significativo se puede considerar como un factor influyente en la variación de la variable dependiente. Un p valor inferior a 0,10 es límite utilizado en

nuestro estudio para considerar una variable significativa, también consideramos un p valor de 0,05 y un p valor de 0,01 el cual sería un coeficiente muy significativo. Es importante, tener en cuenta que un coeficiente significativo al 0,01, también lo es al 0,05 y al 0,10, no obstante, una variable significativa al 0,10, no lo es al 0,05 ni al 0,01.
Podemos encontrarnos que nos salgan muy pocos parámetros significativos individualmente en nuestro modelo, aunque si tenemos al menos un coeficiente significativo individualmente, el modelo será significativo globalmente. No obstante, cuanto mayor sea el número de coeficientes significativos individualmente, mejor será el modelo. También nos podemos encontrar que un coeficiente sea significativo para la clase 'Bajo' u otra clase y no lo sea para el resto de las clases. Por otro lado, que un coeficiente sea no significativo indica que si se mantienen las otras variables constantes la variable con coeficiente no significativo no se puede considerar como un factor influyente en la variación de la variable dependiente. Si observamos los resultados obtenidos, de cada variable explicativa encontramos algún coeficiente significativo individualmente en alguna de las tres clases, por lo que podemos considerar que nuestro modelo también es significativo globalmente. Por ejemplo, podemos observar que la variable Combustible es significativa al 1% en la clase Bajo, significativa al 5% en la clase Medio y no es significativa en la clase Alto, por lo que nos está indicando que esta variable sólo es influyente en la clase Bajo y Medio y no en la clase Alto.

A continuación, se muestran los resultados del modelo en términos de probabilidades, obtenidos en R para los 6 primeros individuos de nuestra base de datos, habiendo utilizado los individuos de la base de datos utilizada para estimar el modelo:

Cliente	Sin Coste	Bajo	Medio	Alto
1	97,84%	0,98%	0,92%	0,25%
2	97,84%	0,98%	0,92%	0,25%
3	97,77%	0,96%	0,99%	0,27%
4	97,77%	0,96%	0,99%	0,27%
5	97,55%	1,29%	0,91%	0,26%
6	97,55%	1,29%	0,91%	0,26%

Tabla 7: Resultados del modelo en términos de probabilidades. *Fuente: Elaboración propia*

Para verificar que el cálculo se ha hecho correctamente, realizaremos un ejemplo con el individuo 1 de nuestra base de datos, el cual tiene las siguientes características:

- Potencia: 'g'
- Antigüedad vehículo: 2
- Edad Conductor: 46
- Marca: F (Japonensas (excepto Nissan) y Coreanas)
- Combustible: Diesel
- Región: R72
- Densidad de habitantes: 0,076 (en miles)

Aplicamos los coeficientes β de las variables explicativas a las variables explicativas X del individuo 1 y obtenemos para cada clase:

$\boldsymbol{Z_{Bajo,1}}$ = -3,894 + 0,046 * 1 – 0,010 * 2 + 0,005 * 46 - 0,838 * 1 - 0,137 * 1 + 0,013 * 0,076 = - 4,612
$\boldsymbol{Z_{Medio,1}}$ = -3,903 + 0,085*1 – 0,001*2 - 0,011*46 - 0,293*1 – 0,067*1 + 0,019*0,076 = - 4,685

$\boldsymbol{Z_{Alto,1}}$ = -4,871+ 0,049*1 - 0,008*2 - 0,020*46 - 0,271*1 + 0,048*1 + 0,000*0,076 = -5,981

En cada caso, los números negativos nos dicen que el individuo 1 tenía más probabilidades de caer en la categoría de referencia 'Sin Coste'. A partir de estos números, podemos calcular que para el individuo 1:

P(Y_1 =Sin Coste) = $\frac{1}{1+exp(-4,612)+exp(-4,685)+exp(-5,981)}$= 0,9784

P(Y_1=Bajo) = $\frac{exp(-4,612)}{1+exp(-4,612)+exp(-4,685)+exp(-5,981)}$= 0,0098

P(Y_1 =Medio) = $\frac{exp(-4,685)}{1+exp(-4,612)+exp(-4,685)+exp(-5,981)}$= 0,0092

P(Y_1 =Alto) = $\frac{exp(-5,981)}{1+exp(-4,612)+exp(-4,685)+exp(-5,981)}$= 0,0025

El individuo 1 tiene un 97,84% de posibilidades de tener un siniestro de la clase 'Sin Coste', un 0,98% de tener un siniestro de la clase 'Bajo', un 0,92% de tener un siniestro de la clase 'Medio' y un 0,25% de posibilidades de tener un siniestro de la clase 'Alto'. Verificamos que estos resultados nos coinciden con los obtenidos tras aplicar el modelo en R.

4.2. Resultados Regresión log-Normal

En esta sección mostraremos los resultados de la modelización de la siguiente expresión $E(Y|C = c, X = x)$. La esperanza del coste condicionada a cada clase de C y a las características de cada individuo.

El enfoque utilizado en el trabajo nos permite modelizar cada tramo de coste según el modelo que mejor ajuste en cada caso. Además, sólo modelizaremos la parte de los siniestros con coste. Como la mayoría de los clientes no tienen siniestros dentro del año, conseguiremos una reducción del tiempo de cálculo, ya que omitiremos los datos de los clientes sin siniestros para modelar las esperanzas condicionales, sin que esto conlleve ninguna pérdida de información.

4.2.1. Regresión Log-Normal para la clase Bajo

A continuación, mostramos los coeficientes (μ_j) obtenidos por el modelo de regresión Log-Normal para la variable dependiente C=Bajo:

		Bajo	
		Coeficientes	**P value**
Referencia		6,513	0,000
Potencia			
	e	0,002	0,952
	f	-0,017	0,626
	g	-0,010	0,770
	h	0,024	0,622
	i	0,069	0,199
	j	0,000	0,998
	k	-0,025	0,722
	l	-0,066	0,530
	m	-0,116	0,443
	n	-0,381	0,027**
	o	-0,285	0,105
Antigüedad		0,006	0,002***
Edad		0,004	0,000***
Marca			
	Japanese (except Nissan) or Korean	0,298	0,000***
	Mercedes, Chrysler or BMW	-0,130	0,049**
	Opel, General Motors or Ford	-0,079	0,155
	other	-0,107	0,167
	Renault, Nissan or Citroen	-0,052	0,288
	Volkswagen, Audi, Skoda or Seat	-0,045	0,438
Combustible			
	Regular	-0,041	0,062*
Región			
	R23	-0,096	0,358
	R24	-0,087	0,027**
	R25	-0,089	0,184
	R31	-0,003	0,950
	R52	-0,073	0,119
	R53	-0,021	0,650
	R54	-0,035	0,533
	R72	-0,067	0,213
	R74	-0,019	0,849
Densidad		-0,005	0,082*

Tabla 8: Resultados del modelo de distribución log-Normal, del cual la variable dependiente es la categoría del coste total de los siniestros por póliza del grupo 'Bajo'. Los asteriscos indican los coeficientes significativos al 1% (***), al 5% (**) y al 10% (*) en el test de significación individual. *Fuente: Elaboración Propia*

En los resultados anteriores se obtienen los coeficientes y los p valores. Estos coeficientes sí que nos están mostrando elasticidades, es decir, como impacta un aumento de la variable explicativa en una unidad, en la variable dependiente clase Bajo. Por ejemplo,

por cada aumento en una unidad de la variable Densidad de habitantes, el logaritmo del coste agregado por póliza de la clase Bajo disminuye en -0,005, o, por otro lado, para los clientes con un vehículo de la marca Renault, Nissan o Citröen, el logaritmo del coste agregado por póliza de la clase Bajo disminuirá en -0,052.

Por otro lado, analizando la significación de los coeficientes, podemos observar como de cada variable explicativa encontramos algún coeficiente significativo individualmente, por lo que podemos considerar que nuestro modelo también es significativo globalmente. Por ejemplo, podemos observar que la categoría 'n' de la variable Potencia es significativa al 5%, por lo que nos está indicando que esta variable es influyente en la clase Bajo. Por otro lado, el resto de las categorías de la variable Potencia son no significativas. También encontramos las variables Antigüedad del vehículo, edad del conductor y la marca del vehículo Japonesas (excepto Nissan) o coreanas, son significativas al 1%, la marca Mercedes, Chrysler o BMW y la región R24 significativas al 5% y por último el combustible Gasolina y Densidad de habitantes significativos al 10%.

Con este modelo estaríamos obteniendo la esperanza del logaritmo del coste, μ=E(ln($Y|C = Bajo, X = x$)), no obstante, lo que nos interesa modelizar es la esperanza del coste, $E(Y|C = Bajo, X = x)$.

Partiendo de la relación entre la distribución Normal y la Log-Normal:

E(Y) = $exp(\mu + \frac{\sigma^2}{2})$
Var(Y) = $exp(2\mu + \sigma^2)(exp(\sigma^2)-1)$

De este modo conseguimos modelizar la esperanza del coste de los siniestros agregado por póliza, condicionadas a las características o factores de riesgo de cada individuo *j* y a que pertenezca a la clase de 'C', Bajo.

A continuación, se muestran los resultados del modelo en términos de esperanza, obtenidos en R para los 6 primeros individuos, habiendo utilizado los individuos de la base de datos utilizada para estimar el modelo:

Cliente	E(Coste Bajo)
1	1.592,10
2	1.592,10
3	1.566,96
4	1.566,96
5	1.552,76
6	1.552,76

Tabla 9: Resultados del modelo en términos de esperanza condicionada.
Fuente: Elaboración propia

Para verificar que el cálculo se ha hecho correctamente, realizaremos un ejemplo con el individuo 1 de nuestra base de datos, el cual tiene las siguientes características:

- Potencia: ‘g’
- Antigüedad vehículo: 2
- Edad Conductor: 46
- Marca: F (Japonensas (excepto Nissan) y Coreanas)
- Combustible: Diesel
- Región: R72
- Densidad de habitantes: 0,076 (en miles)

Aplicamos los coeficientes correspondientes al individuo 1 y obtenemos:

$\mu_{Bajo,1}$ = 6,513 – 0,010*1 + 0,006*2 + 0,004*46 + 0,298*1 – 0,067*1 – 0,005*0,076 = 6,91

Corresponde con el valor esperado del logaritmo del coste bajo para cada individuo de la base de datos para la clase ‘Bajo’.

No obstante, lo que queremos calcular es el valor esperado del coste ‘Bajo’ para cada individuo de la base de datos, por lo que debemos realizar la siguiente transformación, consecuencia de la relación entre la distribución Normal y la Log-Normal:

$$E(Y|C = 'Bajo') = exp(\mu_{Bajo,1} + \frac{\sigma^2}{2}) = \exp(6{,}91 + \frac{0{,}9236689}{2}) = 1.592{,}1 \text{ euros}$$

El parámetro de dispersión σ^2 nos lo da R calculado y es el 0,9236689.

La esperanza del Coste agregado por póliza de la clase Bajo, para el individuo l es de 1.592,1 euros. Verificamos que estos resultados nos coinciden con los obtenidos tras aplicar el modelo en R.

4.2.2. Regresión Log-Normal para la clase Medio

A continuación, mostramos los coeficientes (μ_j) obtenidos por el modelo de regresión Log-Normal para la variable dependiente C=Medio:

		Medio	
		Coeficientes	**P value**
Referencia		8,317	0,000
Potencia			
	e	-0,015	0,527
	f	0,009	0,695
	g	0,038	0,086*
	h	0,023	0,461
	i	0,031	0,371
	j	-0,008	0,827

	k	-0,028	0,519
	l	0,055	0,445
	m	-0,020	0,832
	n	-0,159	0,142
	o	0,017	0,877
Antigüedad		0,000	0,899
Edad		-0,002	0,000***
Marca			
	Japanese (except Nissan) or Korean	-0,080	0,028**
	Mercedes, Chrysler or BMW	-0,054	0,217
	Opel, General Motors or Ford	-0,068	0,068*
	other	0,009	0,867
	Renault, Nissan or Citroen	-0,053	0,101
	Volkswagen, Audi, Skoda or Seat	-0,025	0,506
Combustible			
	Regular	0,016	0,257
Región			
	R23	0,124	0,014**
	R24	-0,005	0,826
	R25	-0,007	0,882
	R31	0,037	0,231
	R52	0,030	0,314
	R53	-0,002	0,939
	R54	0,047	0,183
	R72	0,036	0,237
	R74	0,002	0,972
Densidad		0,000	0,967

Tabla 10: Resultados del modelo de distribución log-Normal, del cual la variable dependiente es la categoría del coste total de los siniestros por póliza del grupo 'Medio'. Los asteriscos indican los coeficientes significativos al 1% (***), al 5% (**) y al 10% (*) en el test de significación individual. *Fuente: Elaboración propia*

Podemos observar, que por cada aumento en una unidad de la variable Edad del Conductor, el logaritmo del coste agregado por póliza de la clase Medio disminuye en -0,002, o por otro lado, para los clientes con un vehículo de marca Japonesa (excepto Nissan) o Coreana, el logaritmo del coste agregado por póliza de la clase Medio, también disminuirá en -0,080.

Por otro lado, podemos observar que la categoría 'g' de la variable Potencia es significativa al 10%, por lo que nos está indicando que esta variable es influyente en la clase Medio. Por otro lado, el resto de las categorías de la variable Potencia son no significativas. La variable marca Opel, General Motors o Ford también es significativa al 10%, mientras que la variable edad del conductor es significativa al 1% y las variables marca Japonesa (excepto Nissan) o Coreanas y la Región R23 son significativas al 5%.

Como hemos comentado anteriormente, con este modelo estaríamos obteniendo la μ=E(ln(Y)), no obstante, lo que nos interesa modelizar es la E(Y).

Partiendo de la relación entre la distribución Normal y la Log-Normal, conseguimos modelizar la esperanza del coste de los siniestros agregado por póliza condicionadas a las características o factores de riesgo de cada individuo *j* y a que pertenezca a la clase de 'C', Medio.

A continuación, se muestran los resultados del modelo en términos de esperanza, obtenidos en R para los 6 primeros individuos, habiendo utilizado los individuos de la base de datos utilizada para estimar el modelo:

Cliente	E(Coste Medio)
1	4.112,37
2	4.112,37
3	4.124,14
4	4.124,14
5	4.124,30
6	4.124,30

Tabla 11: Resultados del modelo en términos de esperanza condicionada.
Fuente: Elaboración propia

Para verificar que el cálculo se ha hecho correctamente, realizaremos un ejemplo con el individuo 1 de nuestra base de datos, el cual tiene las siguientes características:

- Potencia: 'g'
- Antigüedad vehículo: 2
- Edad Conductor: 46
- Marca: F (Japonensas (excepto Nissan) y Coreanas)
- Combustible: Diesel
- Región: R72
- Densidad de habitantes: 0,076 (en miles)

Aplicamos los coeficientes correspondientes al individuo 1 y obtenemos:

$\mu_{Medio,1}$ = 8,317 + 0,038*1 – 0,002*46 – 0,080*1 + 0,036*1 = 8,22

El valor obtenido, corresponde con el valor esperado del logaritmo del coste bajo para cada individuo de la base de datos para la clase 'Medio'.

No obstante, lo que queremos calcular es el valor esperado del coste 'Medio' para cada individuo de la base de datos, por lo que debemos realizar la siguiente transformación, consecuencia de la relación entre la distribución Normal y la Log-Normal:

$$E(Y|C = 'Medio') = exp(\mu_{Medio,1} + \frac{\sigma^2}{2}) = \exp(8{,}22 + \frac{0{,}1910707}{2}) = 4.112{,}37 \text{ euros}$$

El parámetro de dispersión σ^2 nos lo da R calculado y es el 0,1910707.

La esperanza del Coste agregado por póliza de la clase Medio, para el individuo l es de 4.112,37 euros. Verificamos que estos resultados nos coinciden con los obtenidos tras aplicar el modelo en R.

4.2.3. Regresión Log-Normal para la clase Alto

A continuación, mostramos los coeficientes (μ_j) obtenidos por el modelo de regresión Log-Normal para la variable dependiente C=Alto:

		Alto	
		Coeficientes	**P value**
Referencia		10,345	0,000
Potencia			
	e	-0,006	0,957
	f	0,075	0,469
	g	-0,033	0,748
	h	0,021	0,890
	i	-0,040	0,805
	j	-0,084	0,592
	k	0,244	0,312
	l	-0,043	0,868
	m	-0,190	0,681
	n	0,469	0,457
	o	-0,245	0,653
Antigüedad		-0,001	0,898
Edad		-0,004	0,087*
Marc			
	Japanese (except Nissan) or Korean	0,032	0,860
	Mercedes, Chrysler or BMW	0,084	0,679
	Opel, General Motors or Ford	0,018	0,919
	other	-0,064	0,801
	Renault, Nissan or Citroen	-0,018	0,907
	Volkswagen, Audi, Skoda or Seat	-0,133	0,466
Combustible			
	Regular	0,152	0,022**
Región			
	R23	-0,413	0,077*
	R24	0,152	0,175
	R25	-0,103	0,632
	R31	0,019	0,896
	R52	0,123	0,371
	R53	-0,036	0,791
	R54	0,000	0,999

	R72	-0,019	0,892
	R74	-0,148	0,649
Densidad		-0,009	0,239

Tabla 12: Resultados del modelo de distribución log-Normal, del cual la variable dependiente es la categoría del coste total de los siniestros por póliza del grupo 'Alto'. Los asteriscos indican los coeficientes significativos al 1% (***), al 5% (**) y al 10% (*) en el test de significación individual. *Fuente: Elaboración propia*

Podemos observar, que por cada aumento en una unidad de la variable Edad del Conductor, el logaritmo del coste agregado por póliza de la clase Alto disminuye en -0,004, o, por otro lado, para los clientes con un vehículo de Gasolina, el logaritmo del coste agregado por póliza de la clase Alto también aumentará en 0,152.

Por otro lado, podemos observar que sólo nos han salido 3 coeficientes significativos para esta clase, concretamente la variable Edad del conductor al 10%, el tipo de Combustible Gasolina al 5% y la Región R23 al 10%. Esta clase tiene el problema de que no hay suficientes datos para garantizar una estimación precisa, por lo que, aunque nos hayan salido coeficientes significativos, el resultado de esta estimación seguramente será mejorable.

Como hemos comentado anteriormente, con este modelo estaríamos obteniendo la μ=E(ln(Y)), no obstante, lo que nos interesa modelizar es la E(Y).

Partiendo de la relación entre la distribución Normal y la Log-Normal, conseguimos modelizar la esperanza del coste de los siniestros agregado por póliza condicionadas a las características o factores de riesgo de cada individuo *j* y a que pertenezca a la clase de 'C', Alto.

A continuación, se muestran los resultados del modelo en términos de esperanza, obtenidos en R para los 6 primeros individuos, habiendo utilizado los individuos de la base de datos utilizada para estimar el modelo:

Cliente	E(Coste Alto)
1	45.759,05
2	45.759,05
3	61.634,06
4	61.634,06
5	53.699,50
6	53.699,50

Tabla 13: Resultados del modelo en términos de esperanza condicionada.
Fuente: Elaboración propia

Para verificar que el cálculo se ha hecho correctamente, realizaremos un ejemplo con el individuo 1 de nuestra base de datos, el cual tiene las siguientes características:

- Potencia: 'g'
- Antigüedad vehículo: 2
- Edad Conductor: 46
- Marca: F (Japonensas (excepto Nissan) y Coreanas)
- Combustible: Diesel

- Región: R72
- Densidad de habitantes: 0,076 (en miles)

Aplicamos los coeficientes correspondientes al individuo 1 y obtenemos:

$\mu_{Alto,1}$ = 10,345 – 0,033*1 – 0,001*2 – 0,004*46 + 0,032*1 – 0,019*1 – 0,009*0,076 = 10,14

Corresponde con el valor esperado del logaritmo del coste bajo para cada individuo de la base de datos para la clase 'Alto'.

No obstante, lo que queremos calcular es el valor esperado del coste 'Alto' para cada individuo de la base de datos, por lo que debemos realizar la siguiente transformación, consecuencia de la relación entre la distribución Normal y la Log-Normal:

$$E(Y|C = 'Alto') = exp(\mu_{Alto,1} + \frac{\sigma^2}{2}) = \exp(10{,}14 + \frac{1{,}137326}{2}) = 45.759{,}05 \text{ euros}$$

El parámetro de dispersión σ^2 nos lo da R calculado y es el 1,137326.

La esperanza del Coste agregado por póliza de la clase Medio, para el individuo l es de 45.759,05 euros. Verificamos que estos resultados nos coinciden con los obtenidos tras aplicar el modelo en R.

4.3. Resultados combinación de regresiones multivariantes

Ahora ya disponemos de todos los datos para proceder a estimar la prima pura, a través del enfoque aplicado en nuestro trabajo:

$$E(Y|X = x) = \sum_{c=2}^{k} P(C = c|X = x) \cdot E(Y|C = c, X = x)$$

Siguiendo con el ejemplo del individuo 1, vamos a calcular la prima pura resultante de aplicar la combinación de regresiones multivariantes calculadas anteriormente:

E(Y | X= x) = 0,0098· 1.592,1 + 0,0092·4.112,37 + 0,0025·45.759,05 = 167,83 euros

La prima pura resultante para el individuo 1 de la base de datos utilizada es de 167,83 euros.

Si calculamos la prima pura para todos los individuos de nuestra base de datos, esperaríamos obtener un total de 153,3 millones de euros, que es la suma del coste total agregado por póliza. No obstante, tras estimar la prima pura para todos los individuos de nuestra base de datos obtenemos un total de 98,5 millones de euros, existiendo una diferencia de -54,9 millones de euros en nuestra estimación.

Para ver de dónde viene esta diferencia, vamos a comparar el coste real de cada intervalo de C con nuestra estimación. En la Tabla 14 podemos observar estas diferencias:

Coste Bajo real	Coste Bajo Est.	Dif.	Coste Medio real	Coste Medio Est.	Dif.	Coste Alto real	Coste Alto Est.	Dif.	Coste total real	Coste total Est.	Dif.
9.589.110	11.524.782	1.935.672	18.604.617	18.506.402	-98.215	125.149.939	68.415.647	-56.734.292	153.343.666	98.446.831	-54.896.835

Tabla 14: Comparativa entre el resultado estimado y el coste real por póliza. *Fuente: Elaboración propia*

Podemos observar como en el intervalo de 'C' Bajo existe una diferencia de 1,9 millones de euros, en el intervalo Medio una diferencia de 99 mil euros y en el intervalo Alto 56,7 millones de euros. Por lo tanto, el intervalo que mejor ha ajustado nuestro modelo es el intervalo Medio de entre 2.000 y 10.000 euros, seguido por el intervalo Bajo y el que peor hemos ajustado es el intervalo Alto.

Respecto al intervalo Bajo, entre >0 y 2.000 euros, si nos fijamos en su distribución en el Gráfico 13, podemos observar que hay un punto alrededor de los 1.000 euros en el que hay mucha masa de pólizas con este coste, que distorsionan su función de distribución.

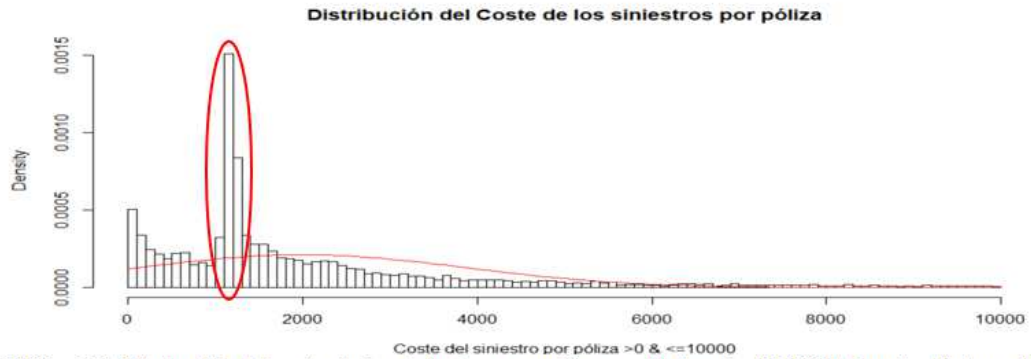


Gráfico 13: Distribución del coste de los siniestros por póliza en el intervalo (0-1000]. Fuente: *Elaboración propia*

Esto puede ser debido a que muchas compañías trabajan por módulos (por ejemplo, en el ramo de autos, el módulo de un taller) y por lo tanto hay muchos siniestros con el mismo coste, por lo que deberíamos tenerlo en cuenta a la hora de modelizar este tramo, para ajustar mejor el coste.

Por otro lado, respecto al intervalo Alto, tiene el problema de que no hay suficientes datos para garantizar una estimación precisa y además tiene valores muy extremos que hacen difícil capturar la cola. Los 16 valores más alejados corresponden a costes de siniestros por póliza por encima del millón de euros y los dos valores más alejados están por encima de los 9 millones de euros.

Estos valores de siniestros tan elevados han sido consecuencia de la transformación que hemos realizado para incorporar la exposición en nuestra base de datos, dividiendo el coste agregado por póliza por la exposición con el objetivo de obtener unos costes comparables entre todas las pólizas. Sin incorporar la exposición, el coste por póliza más elevado era de 2,0 millones de euros. Podríamos haber tenido en cuenta que un siniestro con un coste muy alto es un evento muy poco probable y por lo tanto no dividir el coste por póliza por la exposición para este tramo.

En el box-plot del Gráfico 14 podemos observar como de alejados se encuentran estos valores respecto al resto:

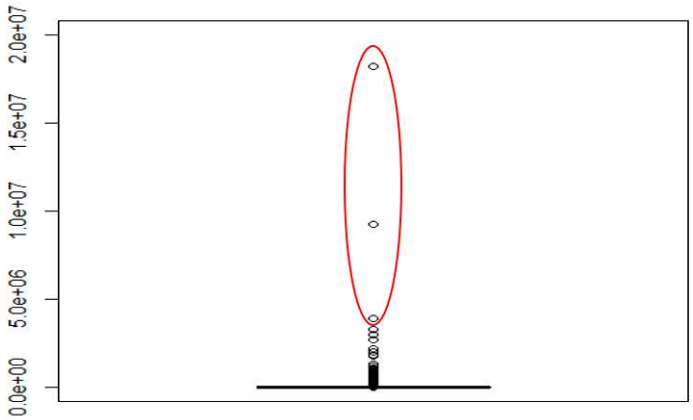


Gráfico 14: Bloxplot variable coste de los siniestros por póliza, corregida. Fuente: Elaboración propia

Por lo tanto, parece que la división y tratamiento que hemos realizado de los datos para este intervalo es mejorable. No obstante, el objetivo de este trabajo es el de demostrar que con la estimación que hemos realizado mediante intervalos del Coste, aun no siendo la mejor estimación posible, es una estimación mejor que la que hubiéramos obtenido sino hubiéramos realizado la estimación por intervalos.

Para demostrarlo, hemos realizado el mismo análisis, pero de la forma tradicional sin intervalos, estimando sólo la probabilidad de 'Sin Coste' y 'Con Coste', mediante la misma combinación de regresiones multivariantes.

Para este nuevo modelo hemos obtenido una prima pura total de 69.107.443 euros, por lo que existe una diferencia de -84,2 millones de euros frente a la suma del coste total agregado por póliza de la base de datos utilizada. En la Tabla 15 podemos ver una comparación entre los resultados de las dos estimaciones realizadas y el coste total por póliza real de la base de datos:

Coste Total real	Coste Estimado enfoque indirecto	Coste Estimado enfoque directo
153.343.666	98.446.831	69.107.443

Tabla 15: Comparativa entre el resultado estimado de forma indirecta, de forma directa y el coste real por póliza. *Fuente: Elaboración propia*

Por lo que queda demostrado que, con el enfoque aplicado en el presente trabajo, se consigue ajustar mejor la prima pura, debido a que, al trocear el coste por intervalos, se consigue que nuestros parámetros vayan variando y vayan dando más margen a la estimación, que si lo modelizamos directamente.

5. Conclusiones

El objetivo principal de este trabajo era modelizar el coste agregado para una cartera del ramo de automóvil a Terceros, para estimar la prima pura, utilizando un nuevo enfoque propuesto por Andreas Christmann en el paper 'An aproach to model complex high-dimensional insurance data', y demostrar que este enfoque tiene ventajas sobre la estimación directa obtenida mediante los métodos tradicionales.

El enfoque consiste en estimar el Coste agregado por póliza condicionado a intervalos de tramos de coste. Después de aplicar este enfoque a la base de datos utilizada, hemos podido observar, que este método es más flexible que otros métodos, ya que permite modelizar cada tramo de coste según el modelo de regresión que mejor ajuste en cada caso, además al modelizar la parte de la esperanza del coste sólo de los siniestros con coste, se consigue una reducción del tiempo de cálculo, ya que la mayoría de clientes no tienen siniestros dentro del año, por lo que sus datos pueden omitirse para modelar las esperanzas condicionales sin ninguna pérdida de información. Por otro lado, hemos podido comprobar que, si comparamos el resultado obtenido troceando la variable coste en 4 intervalos de 'C', contra el resultado obtenido de modelizar directamente la esperanza del coste sin intervalos, el resultado del coste esperado es mejor con el enfoque de Christmann. Entonces podemos concluir que podemos ajustar mejor el coste esperado mediante las esperanzas condicionadas a los diferentes intervalos de 'C' que, modelizando directamente la esperanza del coste agregado, dado que, al trocear por intervalos, nuestros parámetros van a ir variando y nos van a dar más margen, que si lo modelizamos directamente.

En nuestro trabajo, hemos utilizado el modelo Logit Multinimal para la parte de las probabilidades condicionales, no obstante, tras observar los resultados obtenidos, creo que se hubieran podido mejorar, realizando una mejor división de los límites del intervalo de C, entre costes pequeños, grandes o muy grandes, y ajustando un mejor modelo para la clase Alto, la cual tiene el problema de que no hay suficientes datos para garantizar una estimación precisa. Este punto me parece interesante y podría ser un punto de partida para otros trabajos, el conseguir un punto de corte de los datos lo más optima posible, mediante un análisis de los valores extremos de la variable coste agregado por póliza, y poder ajustar en este tramo el mejor modelo, por ejemplo, utilizando el enfoque propuesto en el paper de Max Rydman 'Application of the Peaks-Over-Threshold Method on Insurance Data'. Además, existen otros procedimientos estadísticos de machine learning que son de interés y que podrían utilizarse, como la regresión de Kernel o AdaBoost.

La investigación de diferentes técnicas de estimación de probabilidades condicionales y de las regresiones nos darían como resultado un mejor desempeño del modelo aplicado, que el conseguido.

Por otro lado, aunque la correlación entre variables explicativas está recogida en la estimación de los coeficientes, cuando dos variables explicativas están muy correlacionadas el modelo puede tener un problema de multicolinealidad y en estos casos tendremos problemas en la interpretación de las variables explicativas. Por lo tanto, en estos casos no se podrá hacer caso al valor del coeficiente ni al nivel de significación de las dos variables explicativas. Es decir, no se sabrá cómo en verdad están afectando al valor esperado, aunque, el valor estimado si se esté estimando correctamente. Es decir, no será un problema de ajuste, sino de interpretación de las variables explicativas. Esto

también puede haber sido uno de los motivos por el cual no hemos podido mejorar los resultados del modelo y también me parece un punto interesante a analizar en el futuro, ya que, aunque puede no haber sido el caso de las variables explicativas de nuestro modelo, por ejemplo, la variable edad del conductor y la antigüedad del carnet, por experiencia, suelen ser variables muy correlacionadas, ya que seguro que un conductor de 18 años tendrá poca antigüedad del carnet, entonces si las tuviéramos en nuestro modelo, seguro que no estaríamos interpretando correctamente los coeficientes de las variables si no tuviéramos en cuenta el efecto de la multicolinealidad.

Para acabar, hay que recordar que el principal objetivo de las aseguradoras es determinar la prima de la póliza que ha de pagar el asegurado. Para ello es necesario aplicar modelos que ajusten de la manera más precisa posible la prima pura, y para ello hemos demostrado que la modelización de la suma esperada del importe de los siniestros condicionada por intervalos de coste tiene ventajas sobre la modelización directa del coste esperado.

Bibliografía

Christmann, A. (2004). *An approach to model complex high-dimensional insurance data*. Allgemeines Statistisches Archiv 0, 0–21

Marin-Galiano, M. & Christmann, A. (2004). *Insurance: an R-Program to Model Insurance* Data. Techincal Reports 2004,49. Technische Universität Dortmund, Sonderforschungsbereich 475: Komplexitätsreduktion in multivariaten Datenstrukturen.

Charpentier, A. (2015). *Computational Actuarial Science with R*. Londres: Taylor & Francis Group, an Informa business.

Institute for Digital Research & Education. (13 de Junio de 2014). *Multinomial Logistic Regression | R Data Analysis Examples*. Recuperado de https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/

Williams, R. (2019). *Multinomial Logit Models – Overview*. Recuperado de https://www3.nd.edu/~rwilliam/stats3/Mlogit1.pdf

Paladino, M (5 de Abril de 2017). *Modelos Logit con R*. Recuperado de https://www.institutomora.edu.mx/testU/SitePages/martinpaladino/modelos_logit_con_R.html#modelos-ligit-multinomiales

Boj, E. (1 de Enero de 2006). *Tarificación del seguro del automóvil: Métodos de análisis multivariante*. Recuperado de http://emis.impa.br/EMIS/journals/BEIO/files/BEIOv22n4_AP_EBoj.pdf

Boj, E; Claramunt, M.M. & Costa, T. (2003). *Matemática Actuarial No Vida. Un Enfoque Práctico.* Barcelona: Departamento de Matemática Económica, Financiera y Actuarial, División de Ciencias Jurídicas, Económicas y Sociales, Universidad de Barcelona

Alemany, R; Ayuso, M. & Bolancé, C. (Curso 2019/2020). *Modelos lineales generalizados (tarificación a priori).* Asignatura Modelos estadísticos aplicados. Barcelona: Departamento de Econometría, Estadística y Economía Aplicada, Unicersidad de Barcelona

Ayuso,M & Bolancé,C (Curso 2019/2020). *Manuales del curso de Econometría Actuarial y Análisis de la supervivencia en seguros*. Barcelona: Departamento de Econometría, Estadística y Economía Aplicada, Universidad de Barcelona

Rydman, M. (27 de Junio de 2018). *Application of the Peaks-Over-Threshold Method on Insurance Data.* Suecia: Department of Mathematics Uppsala University

James, G; Witten, D; Hastie, T & Tibshirani, R. (24 de Junio de 2013). *An introduction to Statistical Learning. With Applications in R*

Hastie, T; Tibshirani, R; Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*

ICEA – Servicio de Estadísticas y estudios del sector seguros en España. (2020) *Información del seguro*. Recuperado de https://www.icea.es/es-ES

De Pablos, J. (26 de Septiembre de 2016). *Guía breve sobre la forma de citar y referenciar en ciencias sociales.* Granada: Facultad de Ciencias Políticas y Sociología, Universidad de Granada

Anexo

1.Gràficos sobre la distribución del coste agregado por póliza, según los intervalos de 'C' definidos:

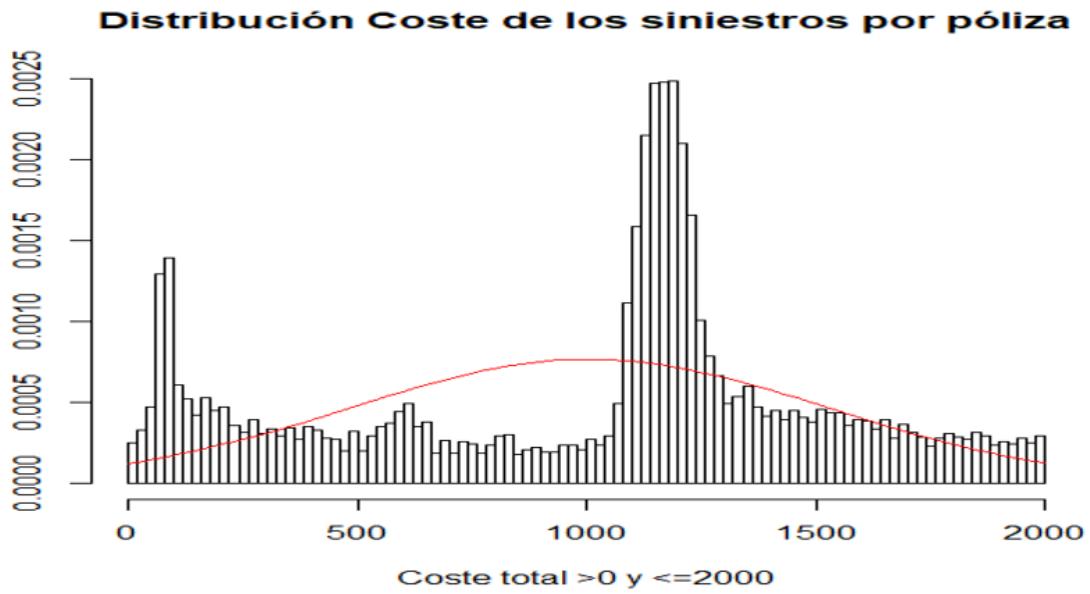


Gráfico 15: Distribución del coste de los siniestros por póliza en el intervalo (0-2.000]. Fuente: Elaboración propia

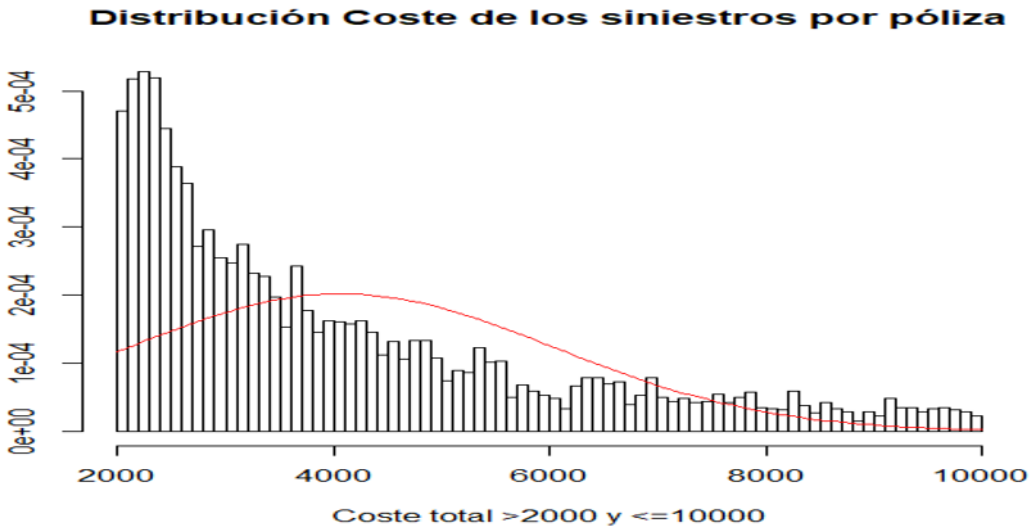


Gráfico 16: Distribución del coste de los siniestros por póliza en el intervalo (2.000-10.000]. Fuente: Elaboración propia

Gráfico 17: Distribución del coste de los siniestros por póliza en el intervalo (>10.000]. Fuente: Elaboración propia

2.Código R:

Instalación de los paquetes necesarios:

```
install.packages("CASdatasets")
library(CASdatasets)
data(freMTPLfreq)
data(freMTPLsev)

install.packages("dplyr")
library(dplyr)

install.packages("ggplot2")
library(ggplot2)

install.packages("gmodels")
library(gmodels)

install.packages("Hmisc")
library(Hmisc)

install.packages("ggthemes")
library(ggthemes)

install.packages("frequency")
library(frequency)

install.packages("foreign")
library(foreign)

install.packages("nnet")
```

```
library(nnet)

install.packages("ggplot2")
library(ggplot2)

install.packages("reshape2")
library(reshape2)

install.packages("nnet")
library(nnet)
```

Preparación y tratamiento de la base de datos:

```
setwd("C:/Users/MATEOARD/Desktop/TFM")
datos<-read.csv("davidMOTORv2.CSV", header=TRUE,sep=";" )

head(datos)
summary(datos)
str(datos)

#valores missing, les asignamos un 0:
datos$claimamount_2 = ifelse(datos$ClaimNb==0, 0, datos$claimamount)

#creamos variable para incorporar la exposición:
datos$claimamount_corregida<-datos$claimamount_2/datos$Exposure

#dividimos la variable density en 1000:
datos$densityM<-datos$Density/1000

glimpse(datos)
table(datos$Power)
CrossTable(datos$Power)
hist(datos$claimamount_corregida)
boxplot(datos$claimamount_corregida, ylim=c(0, 20000000))
describe(datos)

#Creamos la variable Rango de coste de siniestro por póliza:

datos$claimamount_rangos<-0
datos$claimamount_rangos[datos$claimamount_corregida==0]<-"Sin Coste"
datos$claimamount_rangos[datos$claimamount_corregida>0                    &
datos$claimamount_corregida<=2000]<-"Bajo"
datos$claimamount_rangos[datos$claimamount_corregida>2000                 &
datos$claimamount_corregida<=10000]<-"Medio"
datos$claimamount_rangos[datos$claimamount_corregida>10000]<-"Alto"
wtd.table(datos$claimamount_rangos)

datos$claimamount_rangos = factor(datos$claimamount_rangos,
                    levels=c("Sin Coste","Bajo", "Medio", "Alto"))
describe(datos)
```

```
table(datos$claimamount_rangos)

#Creamos otra variable Rango de coste de siniestros por póliza:

datos$claimamount_rangos2<-0
datos$claimamount_rangos2[datos$claimamount_corregida==0]<-"Sin Coste"
datos$claimamount_rangos2[datos$claimamount_corregida>0]<-"Con Coste"
wtd.table(datos$claimamount_rangos2)

datos$claimamount_rangos2 = factor(datos$claimamount_rangos2,
                    levels=c("Sin Coste","Con Coste"))

table(datos$claimamount_rangos2)

#Creamos la variable Rango edad conductor:

Edades_Rango<-case_when(
  datos$DriverAge<31 ~ "(18-30]",
  datos$DriverAge<46 ~ "(30-45]",
  datos$DriverAge<56 ~ "(45-55]",
  datos$DriverAge<66 ~ "(55-65]",
  TRUE ~ ">65")

table(Edades_Rango)

par(mfrow=c(2,2))
xnames<-names(tapply(datos$ClaimNb, Edades_Rango,mean))
plot(tapply(datos$ClaimNb, Edades_Rango,mean), main="Frecuencia del conjunto de
siniestros por póliza y Rangos de edad del Conductor",xlab="Rangos de edad del
Conductor", ylab="Frecuencia media de siniestros por póliza" , type="b", xaxt="n")
axis(1, at=1:length(xnames), labels=xnames)
xnames<-names(tapply(datos$claimamount_corregida, Edades_Rango,mean))
plot(tapply(datos$claimamount_corregida, Edades_Rango,mean),main="Coste medio
del conjunto de siniestro por póliza y Rangos de edad del Conductor",xlab="Rangos de
edad del Conductor",ylab="Coste Medio de siniestro por póliza" , type="b", xaxt="n")
axis(1, at=1:length(xnames), labels=xnames)

#coeficiente de correlación spearman:
cor(datos$ClaimNb, datos$DriverAge, method="spearman")
cor(datos$claimamount_corregida, datos$DriverAge, method="spearman")

#Creamos la variable rango antigüedad vehículo:

Antiguedad_Rango<-case_when(
  datos$CarAge<1 ~ "0",
  datos$CarAge<2 ~ "1",
  datos$CarAge<3 ~ "2",
  datos$CarAge<5 ~ "(2-5]",
  datos$CarAge<10 ~ "(5-10]",
```

```
  TRUE ~ ">10")

table(Antiguedad_Rango)

Antiguedad_Rango = factor(Antiguedad_Rango,
                          levels=c("0","1", "2", "(2-5]","(5-10]",">10" ))

par(mfrow=c(2,2))
xnames<-names(tapply(datos$ClaimNb, Antiguedad_Rango,mean))
plot(tapply(datos$ClaimNb, Antiguedad_Rango,mean), main="Frecuencia del conjunto de siniestros por póliza y Rangos de Antigüedad del vehículo",xlab="Rangos de Antigüedad del vehículo", ylab="Frecuencia media de siniestros por póliza" , type="b", xaxt="n")
axis(1, at=1:length(xnames), labels=xnames)
xnames<-names(tapply(datos$claimamount_corregida, Antiguedad_Rango,mean))
plot(tapply(datos$claimamount_corregida, Antiguedad_Rango,mean),main="Coste medio del conjunto de siniestro por póliza y Rangos de edad del Conductor",xlab="Rangos de Antigüedad del vehículo",ylab="Coste Medio de siniestro por póliza" , type="b", xaxt="n")
axis(1, at=1:length(xnames), labels=xnames)

table(Antiguedad_Rango)

#coeficiente de correlación spearman:
cor(datos$ClaimNb, datos$CarAge, method="spearman")
cor(datos$claimamount_corregida, datos$CarAge, method="spearman")

#Creamos la variable rango Densidad habitantes:

Densidad_Rango<-case_when(
  datos$Density<200 ~ "(0-200]",
  datos$Density<500 ~ "(200-500]",
  datos$Density<1000 ~ "(500-1000]",
  datos$Density<2000 ~ "(1000-2000]",
  TRUE ~ ">2000")

Densidad_Rango = factor(Densidad_Rango,
                        levels=c("(0-200]","(200-500]","(500-1000]","(1000-2000]",">2000"))

par(mfrow=c(2,2))
xnames<-names(tapply(datos$ClaimNb, Densidad_Rango,mean))
plot(tapply(datos$ClaimNb, Densidad_Rango,mean), main="Frecuencia del conjunto de siniestros por póliza y Rangos de Densidad de habitantes",xlab="Rangos de Densidad de habitantes", ylab="Frecuencia media de siniestros por póliza" , type="b", xaxt="n")
axis(1, at=1:length(xnames), labels=xnames)
xnames<-names(tapply(datos$claimamount_corregida, Densidad_Rango,mean))
plot(tapply(datos$claimamount_corregida, Densidad_Rango,mean),main="Coste medio del conjunto de siniestro por póliza y Rangos de edad del Conductor",xlab="Rangos de Densidad de habitantes",ylab="Coste Medio de siniestro por póliza" , type="b", xaxt="n")
```

```
axis(1, at=1:length(xnames), labels=xnames)

#coeficiente de correlación spearman:
cor(datos$ClaimNb, datos$Density, method="spearman")
cor(datos$claimamount_corregida, datos$Density, method="spearman")

#Realizamos un análisis descriptivo de los diferentes factores de riesgo:

mean(datos$claimamount_corregida)
sd(datos$claimamount_corregida)
var(datos$claimamount_corregida)

mean(datos$claimamount_corregida[datos$Power=='d'])
mean(datos$claimamount_corregida[datos$Power=='e'])
mean(datos$claimamount_corregida[datos$Power=='f'])
mean(datos$claimamount_corregida[datos$Power=='g'])
mean(datos$claimamount_corregida[datos$Power=='h'])
mean(datos$claimamount_corregida[datos$Power=='i'])
mean(datos$claimamount_corregida[datos$Power=='j'])
mean(datos$claimamount_corregida[datos$Power=='k'])
mean(datos$claimamount_corregida[datos$Power=='l'])
mean(datos$claimamount_corregida[datos$Power=='m'])
mean(datos$claimamount_corregida[datos$Power=='n'])
mean(datos$claimamount_corregida[datos$Power=='o'])

sum(datos$claimamount_corregida[datos$Power=='d'])
sum(datos$claimamount_corregida[datos$Power=='e'])
sum(datos$claimamount_corregida[datos$Power=='f'])
sum(datos$claimamount_corregida[datos$Power=='g'])
sum(datos$claimamount_corregida[datos$Power=='h'])
sum(datos$claimamount_corregida[datos$Power=='i'])
sum(datos$claimamount_corregida[datos$Power=='j'])
sum(datos$claimamount_corregida[datos$Power=='k'])
sum(datos$claimamount_corregida[datos$Power=='l'])
sum(datos$claimamount_corregida[datos$Power=='m'])
sum(datos$claimamount_corregida[datos$Power=='n'])
sum(datos$claimamount_corregida[datos$Power=='o'])

sum(datos$ClaimNb[datos$Power=='d'])
sum(datos$ClaimNb[datos$Power=='e'])
sum(datos$ClaimNb[datos$Power=='f'])
sum(datos$ClaimNb[datos$Power=='g'])
sum(datos$ClaimNb[datos$Power=='h'])
sum(datos$ClaimNb[datos$Power=='i'])
sum(datos$ClaimNb[datos$Power=='j'])
sum(datos$ClaimNb[datos$Power=='k'])
sum(datos$ClaimNb[datos$Power=='l'])
sum(datos$ClaimNb[datos$Power=='m'])
sum(datos$ClaimNb[datos$Power=='n'])
sum(datos$ClaimNb[datos$Power=='o'])
```

```
sd(datos$claimamount_corregida[datos$Power=='d'])
sd(datos$claimamount_corregida[datos$Power=='e'])
sd(datos$claimamount_corregida[datos$Power=='f'])
sd(datos$claimamount_corregida[datos$Power=='g'])
sd(datos$claimamount_corregida[datos$Power=='h'])
sd(datos$claimamount_corregida[datos$Power=='i'])
sd(datos$claimamount_corregida[datos$Power=='j'])
sd(datos$claimamount_corregida[datos$Power=='k'])
sd(datos$claimamount_corregida[datos$Power=='l'])
sd(datos$claimamount_corregida[datos$Power=='m'])
sd(datos$claimamount_corregida[datos$Power=='n'])
sd(datos$claimamount_corregida[datos$Power=='o'])

var(datos$claimamount_corregida[datos$Power=='d'])
var(datos$claimamount_corregida[datos$Power=='e'])
var(datos$claimamount_corregida[datos$Power=='f'])
var(datos$claimamount_corregida[datos$Power=='g'])
var(datos$claimamount_corregida[datos$Power=='h'])
var(datos$claimamount_corregida[datos$Power=='i'])
var(datos$claimamount_corregida[datos$Power=='j'])
var(datos$claimamount_corregida[datos$Power=='k'])
var(datos$claimamount_corregida[datos$Power=='l'])
var(datos$claimamount_corregida[datos$Power=='m'])
var(datos$claimamount_corregida[datos$Power=='n'])
var(datos$claimamount_corregida[datos$Power=='o'])

mean(datos$claimamount_corregida[datos$Brand=='Fiat'])
mean(datos$claimamount_corregida[datos$Brand=='Volkswagen, Audi, Skoda or Seat'])
mean(datos$claimamount_corregida[datos$Brand=='Renault, Nissan or Citroen'])
mean(datos$claimamount_corregida[datos$Brand=='other'])
mean(datos$claimamount_corregida[datos$Brand=='Opel, General Motors or Ford'])
mean(datos$claimamount_corregida[datos$Brand=='Mercedes, Chrysler or BMW'])
mean(datos$claimamount_corregida[datos$Brand=='Japanese (except Nissan) or Korean'])

var(datos$claimamount_corregida[datos$Brand=='Fiat'])
var(datos$claimamount_corregida[datos$Brand=='Volkswagen, Audi, Skoda or Seat'])
var(datos$claimamount_corregida[datos$Brand=='Renault, Nissan or Citroen'])
var(datos$claimamount_corregida[datos$Brand=='other'])
var(datos$claimamount_corregida[datos$Brand=='Opel, General Motors or Ford'])
var(datos$claimamount_corregida[datos$Brand=='Mercedes, Chrysler or BMW'])
var(datos$claimamount_corregida[datos$Brand=='Japanese (except Nissan) or Korean'])

sd(datos$claimamount_corregida[datos$Brand=='Fiat'])
sd(datos$claimamount_corregida[datos$Brand=='Volkswagen, Audi, Skoda or Seat'])
sd(datos$claimamount_corregida[datos$Brand=='Renault, Nissan or Citroen'])
sd(datos$claimamount_corregida[datos$Brand=='other'])
sd(datos$claimamount_corregida[datos$Brand=='Opel, General Motors or Ford'])
```

```
sd(datos$claimamount_corregida[datos$Brand=='Mercedes, Chrysler or BMW'])
sd(datos$claimamount_corregida[datos$Brand=='Japanese (except Nissan) or Korean'])

sum(datos$claimamount_corregida[datos$Brand=='Fiat'])
sum(datos$claimamount_corregida[datos$Brand=='Volkswagen, Audi, Skoda or Seat'])
sum(datos$claimamount_corregida[datos$Brand=='Renault, Nissan or Citroen'])
sum(datos$claimamount_corregida[datos$Brand=='other'])
sum(datos$claimamount_corregida[datos$Brand=='Opel, General Motors or Ford'])
sum(datos$claimamount_corregida[datos$Brand=='Mercedes, Chrysler or BMW'])
sum(datos$claimamount_corregida[datos$Brand=='Japanese (except Nissan) or Korean'])

sum(datos$ClaimNb[datos$Brand=='Fiat'])
sum(datos$ClaimNb[datos$Brand=='Volkswagen, Audi, Skoda or Seat'])
sum(datos$ClaimNb[datos$Brand=='Renault, Nissan or Citroen'])
sum(datos$ClaimNb[datos$Brand=='other'])
sum(datos$ClaimNb[datos$Brand=='Opel, General Motors or Ford'])
sum(datos$ClaimNb[datos$Brand=='Mercedes, Chrysler or BMW'])
sum(datos$ClaimNb[datos$Brand=='Japanese (except Nissan) or Korean'])

mean(datos$claimamount_corregida[datos$Gas=='Diesel'])
mean(datos$claimamount_corregida[datos$Gas=='Regular'])

sum(datos$claimamount_corregida[datos$Gas=='Diesel'])
sum(datos$claimamount_corregida[datos$Gas=='Regular'])

sum(datos$ClaimNb[datos$Gas=='Diesel'])
sum(datos$ClaimNb[datos$Gas=='Regular'])

var(datos$claimamount_corregida[datos$Gas=='Diesel'])
var(datos$claimamount_corregida[datos$Gas=='Regular'])

sd(datos$claimamount_corregida[datos$Gas=='Diesel'])
sd(datos$claimamount_corregida[datos$Gas=='Regular'])

mean(datos$claimamount_corregida[datos$Region=='R11'])
mean(datos$claimamount_corregida[datos$Region=='R23'])
mean(datos$claimamount_corregida[datos$Region=='R24'])
mean(datos$claimamount_corregida[datos$Region=='R25'])
mean(datos$claimamount_corregida[datos$Region=='R31'])
mean(datos$claimamount_corregida[datos$Region=='R52'])
mean(datos$claimamount_corregida[datos$Region=='R53'])
mean(datos$claimamount_corregida[datos$Region=='R54'])
mean(datos$claimamount_corregida[datos$Region=='R72'])
mean(datos$claimamount_corregida[datos$Region=='R74'])

sum(datos$claimamount_corregida[datos$Region=='R11'])
sum(datos$claimamount_corregida[datos$Region=='R23'])
sum(datos$claimamount_corregida[datos$Region=='R24'])
sum(datos$claimamount_corregida[datos$Region=='R25'])
```

```
sum(datos$claimamount_corregida[datos$Region=='R31'])
sum(datos$claimamount_corregida[datos$Region=='R52'])
sum(datos$claimamount_corregida[datos$Region=='R53'])
sum(datos$claimamount_corregida[datos$Region=='R54'])
sum(datos$claimamount_corregida[datos$Region=='R72'])
sum(datos$claimamount_corregida[datos$Region=='R74'])

sum(datos$ClaimNb[datos$Region=='R11'])
sum(datos$ClaimNb[datos$Region=='R23'])
sum(datos$ClaimNb[datos$Region=='R24'])
sum(datos$ClaimNb[datos$Region=='R25'])
sum(datos$ClaimNb[datos$Region=='R31'])
sum(datos$ClaimNb[datos$Region=='R52'])
sum(datos$ClaimNb[datos$Region=='R53'])
sum(datos$ClaimNb[datos$Region=='R54'])
sum(datos$ClaimNb[datos$Region=='R72'])
sum(datos$ClaimNb[datos$Region=='R74'])

var(datos$claimamount_corregida[datos$Region=='R11'])
var(datos$claimamount_corregida[datos$Region=='R23'])
var(datos$claimamount_corregida[datos$Region=='R24'])
var(datos$claimamount_corregida[datos$Region=='R25'])
var(datos$claimamount_corregida[datos$Region=='R31'])
var(datos$claimamount_corregida[datos$Region=='R52'])
var(datos$claimamount_corregida[datos$Region=='R53'])
var(datos$claimamount_corregida[datos$Region=='R54'])
var(datos$claimamount_corregida[datos$Region=='R72'])
var(datos$claimamount_corregida[datos$Region=='R74'])

sd(datos$claimamount_corregida[datos$Region=='R11'])
sd(datos$claimamount_corregida[datos$Region=='R23'])
sd(datos$claimamount_corregida[datos$Region=='R24'])
sd(datos$claimamount_corregida[datos$Region=='R25'])
sd(datos$claimamount_corregida[datos$Region=='R31'])
sd(datos$claimamount_corregida[datos$Region=='R52'])
sd(datos$claimamount_corregida[datos$Region=='R53'])
sd(datos$claimamount_corregida[datos$Region=='R54'])
sd(datos$claimamount_corregida[datos$Region=='R72'])
sd(datos$claimamount_corregida[datos$Region=='R74'])

mean(datos$claimamount_corregida[datos$claimamount_rangos=="Sin Coste"])
mean(datos$claimamount_corregida[datos$claimamount_rangos=="Bajo"])
mean(datos$claimamount_corregida[datos$claimamount_rangos=="Medio"])
mean(datos$claimamount_corregida[datos$claimamount_rangos=="Alto"])

sum(datos$claimamount_corregida[datos$claimamount_rangos=="Sin Coste"])
sum(datos$claimamount_corregida[datos$claimamount_rangos=="Bajo"])
sum(datos$claimamount_corregida[datos$claimamount_rangos=="Medio"])
sum(datos$claimamount_corregida[datos$claimamount_rangos=="Alto"])
```

```
median(datos$claimamount_corregida[datos$claimamount_rangos=="Sin Coste"])
median(datos$claimamount_corregida[datos$claimamount_rangos=="Bajo"])
median(datos$claimamount_corregida[datos$claimamount_rangos=="Medio"])
median(datos$claimamount_corregida[datos$claimamount_rangos=="Alto"])

sd(datos$claimamount_corregida[datos$claimamount_rangos=="Sin Coste"])
sd(datos$claimamount_corregida[datos$claimamount_rangos=="Bajo"])
sd(datos$claimamount_corregida[datos$claimamount_rangos=="Medio"])
sd(datos$claimamount_corregida[datos$claimamount_rangos=="Alto"])

var(datos$claimamount_corregida[datos$claimamount_rangos=="Sin Coste"])
var(datos$claimamount_corregida[datos$claimamount_rangos=="Bajo"])
var(datos$claimamount_corregida[datos$claimamount_rangos=="Medio"])
var(datos$claimamount_corregida[datos$claimamount_rangos=="Alto"])

#realizamos gráficos ilustrativos:
hist(datos$claimamount_corregida[datos$claimamount_corregida>0 & datos$claimamount_corregida<=10000], breaks= 100, main= "Distribución del Coste de los siniestros por póliza", xlab="Coste del siniestro por póliza >0 & <=10000", freq=FALSE)
curve(dnorm(x,
mean=mean(datos$claimamount_corregida[datos$claimamount_corregida>0 & datos$claimamount_corregida<=10000]),
sd=sd(datos$claimamount_corregida[datos$claimamount_corregida>0 & datos$claimamount_corregida<=10000])), add= TRUE, col="red")

par(mfrow=c(2,2))
plot(tapply(datos$ClaimNb, datos$CarAge,mean),main="Frecuencia de siniestros por Antigüedad del vehículo",xlab="Antigüedad del vehículo", ylab="Frecuencia de siniestros", type = "b")
plot(tapply(datos$claimamount_corregida, datos$CarAge,mean),main="Coste medio de siniestro por Antigüedad del Vehículo",xlab="Antigüedad del vehículo",ylab="Coste Medio de siniestro", type = "b")

par(mfrow=c(2,2))
plot(tapply(datos$ClaimNb, datos$DriverAge,mean),main="Frecuencia de siniestros por Edad del Conductor",xlab="Edad del Conductor", ylab="Frecuencia de siniestros", type = "b")
plot(tapply(datos$claimamount_corregida, datos$DriverAge,mean),main="Coste medio de siniestro por Edad del Conductor",xlab="Edad del Conductor",ylab="Coste Medio de siniestro", type = "b")

par(mfrow=c(2,2))
plot(tapply(datos$ClaimNb, datos$Density,mean),main="Frecuencia de siniestros por Densidad de Habitantes",xlab="Densidad de Habitantes", ylab="Frecuencia de siniestros", type = "b")
plot(tapply(datos$claimamount_corregida, datos$Density,mean),main="Coste medio de siniestro por Densidad de Habitantes",xlab="Densidad de Habitantes",ylab="Coste Medio de siniestro", type = "b")
```

Construcción del Logit Multinomial:

```
datos$datos$claimamount_ML <- relevel(datos$claimamount_rangos, ref = "Sin Coste")
test2 <- multinom(datos$claimamount_ML ~ factor(Power) + CarAge + DriverAge + factor(Brand) + factor(Gas) + factor(Region) + densityM, data = datos,maxit=1000)
summary(test2)
z<-summary(test2)$coefficients/summary(test2)$standard.errors
z
p<-(1 - pnorm(abs(z), 0, 1)) * 2
p
exp(coef(test2))
head(pp2 <- fitted(test2))
pp2 <- fitted(test2)
```

Construcción del modelo Log-Normal:

```
#dividimos la BBD en los 4 grupos de coste total de los siniestros por póliza

Sin_Coste<-subset(datos, claimamount_rangos=='Sin Coste')
Bajo<-subset(datos, claimamount_rangos=='Bajo')
Medio<-subset(datos, claimamount_rangos=='Medio')
Alto<-subset(datos, claimamount_rangos=='Alto')
```

#Para cada conjunto de datos (‘Bajo’,’Medio’,’Alto’), ajustamos un modelo de regresión en el que el coste es la variable dependiente. Como variables explicativas se utilizan las mismas que en el Logit Multinomial:

```
Bajo_glm2<-glm(log(claimamount_corregida) ~ Power + CarAge + DriverAge + Brand + Gas + Region + densityM, data=Bajo, family=gaussian)
Medio_glm2<-glm(log(claimamount_corregida) ~ Power + CarAge + DriverAge + Brand + Gas + Region + densityM, data=Medio, family=gaussian)
Alto_glm2<-glm(log(claimamount_corregida) ~ Power + CarAge + DriverAge + Brand + Gas + Region + densityM, data=Alto, family=gaussian)

summary(Bajo_glm2)
summary(Medio_glm2)
summary(Alto_glm2)

#valor esperado del logaritmo del coste para cada individuo de la base de datos 'datos':
mbajo2<-predict(Bajo_glm2,newdata=datos)
mmedio2<-predict(Medio_glm2,newdata=datos)
malto2<-predict(Alto_glm2,newdata=datos)


#valor esperado del coste para cada individuo de la base de datos 'datos'
costebajo2<-exp(mbajo2+(summary(Bajo_glm2)$dispersion)/2)
costemedio2<-exp(mmedio2+(summary(Medio_glm2)$dispersion)/2)
costealto2<-exp(malto2+(summary(Alto_glm2)$dispersion)/2)
```

Cálculo prima pura:

```
primalength <- length(pp2[,1])*5
primapura<- 1:primalength
matrixprimapura <- matrix(data=primapura,nrow=length(pp2[,1]),ncol=5)
colnames(matrixprimapura)<- c('Sin Coste', 'Bajo', 'Medio', 'Alto', 'Prima Pura')

for(r in 1:nrow(pp2)){
  matrixprimapura[r,1] <- pp2[r,1] * 0
  matrixprimapura[r,2] <- pp2[r,2] * costebajo2[r]
  matrixprimapura[r,3] <- pp2[r,3] * costemedio2[r]
  matrixprimapura[r,4] <- pp2[r,4] * costealto2[r]

  matrixprimapura[r,5] <-
matrixprimapura[r,1]+matrixprimapura[r,2]+matrixprimapura[r,3]+matrixprimapura[r,4
]
}

sumaprimapura<-colSums(matrixprimapura[,5, drop = FALSE])
sumaprimapura

#comparamos los resultados obtenidos en el modelo y los datos reales de la base de datos:
c(sum(datos$claimamount_corregida),sumaprimapura)

#realizamos la modelización mediante el enfoque tradicional:

#Logit Multinomial:
datos$datos$claimamount_ML <- relevel(datos$claimamount_rangos2, ref = "Sin
Coste")
test3 <- multinom(datos$claimamount_ML ~ factor(Power) + CarAge + DriverAge +
factor(Brand) + factor(Gas) + factor(Region) + densityM, data = datos,maxit=1000)
summary(test3)
z<-summary(test3)$coefficients/summary(test3)$standard.errors
z
p<-(1 - pnorm(abs(z), 0, 1)) * 2
p
exp(coef(test3))
head(pp3 <- fitted(test3))
pp3 <- fitted(test3)

#Regresión Log-Normal:
Sin_Coste2<-subset(datos, claimamount_rangos2=='Sin Coste')
Con_Coste<-subset(datos, claimamount_rangos2=='Con Coste')

Con_Coste_glm2<-glm(log(claimamount_corregida) ~ Power + CarAge + DriverAge +
Brand + Gas + Region + densityM, data=Con_Coste, family=gaussian)
summary(Con_Coste_glm2)
ccoste2<-predict(Con_Coste_glm2,newdata=datos)
ccostemedio2<-exp(ccoste2+(summary(Con_Coste_glm2)$dispersion)/2)
primapura3<-pp3[,2] * ccostemedio2
sumaprimapura3<-sum(primapura3)
```

#comparamos los resultados obtenidos en los dos modelos y los datos reales de la base de datos:
c(sum(datos$claimamount_corregida),sumaprimapura, sumaprimapura3)